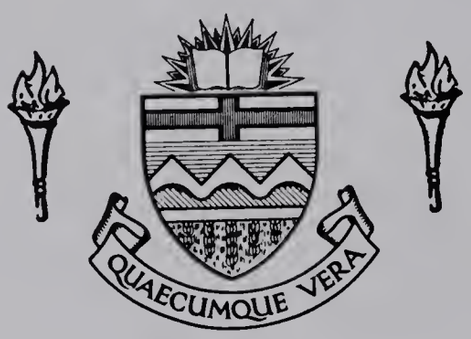


The University of Alberta
Printing Department
Edmonton, Alberta

THE UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR Dolores Siu Kim Lam

TITLE OF THESIS A Study of the Behrens-Fisher Test for the Behrens-Fisher Problem

DEGREE FOR WHICH THESIS WAS PRESENTED M. Sc.

YEAR THIS DEGREE GRANTED Fall, 1974

THE UNIVERSITY OF ALBERTA

A STUDY OF THE BEHRENS-FISHER TEST FOR THE BEHRENS-FISHER PROBLEM

by

© DOLORES SIU KIM LAM

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

OF MASTER OF SCIENCE

DEPARTMENT OF COMPUTING SCIENCE

EDMONTON, ALBERTA

FALL, 1974

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled A Study of the Behrens-Fisher Test for the Behrens-Fisher Problem

submitted by Dolores Siu Kim Lam

in partial fulfilment of the requirements for the degree of Master of Science

To my Mother.

ABSTRACT

This thesis studies the Behrens-Fisher test for the Behrens-Fisher problem. Fisher's hypotheses and his proposed methods of verification of the test are discussed. A sampling study on the actual size of the test based on Fisher's procedures is conducted. Actual sizes for different parameter values are obtained and tabulated. These results are discussed along with those obtained by other investigators. It is shown that the actual sizes as calculated using Fisher's proposed methods are close to the nominal sizes of the test. Furthermore, it is seen that the actual sizes for larger degrees of freedom agree more closely with the nominal sizes than those for smaller degrees of freedom. The test is thus recommended for use because it actually yields actual size close to the size specified by the user.

ACKNOWLEDGEMENTS

I wish to express my sincere thanks and appreciation to Dr. Eric N. West for his guidance and advice throughout the period of my research.

I also wish to thank my husband and my sisters for their constant help and encouragement throughout the period of my study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I. INTRODUCTION

I. Statement of the Problem

Suppose samples of sizes $n_1 + 1$, $n_2 + 1$ are taken separately from two normal populations with true means μ_1, μ_2 and true variances σ_1^2 and σ_2^2. Let the means of the two samples be $\bar{x}_1$ and $\bar{x}_2$ and their variances be $(n_1 + 1)S_1^2$ and $(n_2 + 1)S_2^2$. On the basis of these samples, we want to test the hypotheses that there is no difference between the means of the two populations. That is, we want to test the hypothesis, $H_o : \mu_1 = \mu_2$.

Two cases may occur:

(i) σ_1 and σ_2 are known or equal or in a known ratio.

(ii) σ_1 and σ_2 are not known or not in a known ratio.

For case (i), the problem of the test of hypothesis is easily solved. There are three situations:

(a) σ_1 and σ_2 known.

Now $\bar{x}_1 \sim N(\mu_1, \frac{\sigma_1^2}{n_1 + 1})$, $\bar{x}_2 \sim N(\mu_2, \frac{\sigma_2^2}{n_2 + 1})$.*

Let $d = \bar{x}_1 - \bar{x}_2$. It is easily shown that:

$$E(d) = \mu_1 - \mu_2$$

$$\text{var}(d) = \frac{\sigma_1^2}{n_1 + 1} + \frac{\sigma_2^2}{n_2 + 1}$$

* where $x \sim N(\mu, \sigma^2)$ is read "x has the normal distribution with mean μ and variance σ^2".

If σ_1 and σ_2 are known, the statistic

$$v = \frac{d - \mu_1 + \mu_2}{\left(\frac{\sigma_1^2}{n_1 + 1} + \frac{\sigma_2^2}{n_2 + 1}\right)^{1/2}} \sim N(0,1),$$

and hence the test of hypothesis is easily carried out using the standard normal table.

(b) $\sigma_1^2 = \sigma_2^2 = \sigma^2$ but the value of σ^2 is not known. Use the t test to test the significance of the difference between the two means as follows:

$$\text{Let } S_1^2 = \frac{\sum_{i=1}^{n_1 + 1} (x_{1i} - \bar{x}_1)^2}{n_1 (n_1 + 1)}$$

$$S_2^2 = \frac{\sum_{i=1}^{n_2 + 1} (x_{2i} - \bar{x}_2)^2}{n_2 (n_2 + 1)}$$

We know that

$\frac{n_1(n_1 + 1)S_1^2}{\sigma_1^2}$, $\frac{n_2(n_2 + 1)S_2^2}{\sigma_2^2}$ are independently distributed as $\chi^2_{n_1 df}$ and $\chi^2_{n_2 df}$ respectively. If $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then

$\frac{n_1(n_1 + 1)S_1^2 + n_2(n_2 + 1)S_2^2}{\sigma^2}$ is distributed as χ^2 with $(n_1 + n_2)$ degrees of freedom.

Now

$$\frac{d - (\mu_1 - \mu_2)}{\sigma\left(\frac{1}{n_1 + 1} + \frac{1}{n_2 + 1}\right)^{1/2}} \sim N(0,1)$$

Let

$$t = \left\{ \frac{d - (\mu_1 - \mu_2)}{\sigma\left(\frac{1}{n_1 + 1} + \frac{1}{n_2 + 1}\right)^{1/2}} \right\} \Bigg/ \left\{ \frac{[n_1(n_1 + 1)S_1^2 + n_2(n_2 + 1)S_2^2]}{\cdot\ \sigma^2 (n_1 + n_2)} \right\}^{1/2}$$

$$= \{d - (\mu_1 - \mu_2)\} \Bigg/ \left\{ \frac{n_1(n_1 + 1)S_1^2 + n_2(n_2 + 1)S_2^2}{n_1 + n_2} \right.$$

$$\left. \times \left(\frac{1}{n_1 + 1} + \frac{1}{n_2 + 1}\right) \right\}^{1/2} \qquad (1.1)$$

t which does not now involve the unknown σ^2 is then distributed with Student's t distribution with $n_1 + n_2$ degrees of freedom.

(c) If σ_1^2 and σ_2^2 are in a known ratio, say $\theta = \sigma_1^2/\sigma_2^2$ is known, the t statistic can still be used to perform the test. This is shown as follows :

By analogy with (1.1), if $\sigma_1^2 \neq \sigma_2^2$, we have

$$t = \frac{d - \mu_1 + \mu_2}{\left\{\frac{\sigma_1^2}{n_1 + 1} + \frac{\sigma_2^2}{n_2 + 1}\right\}^{1/2}} \left\{ \frac{\frac{n_1(n_1 + 1)S_1^2}{\sigma_1^2} + \frac{n_2(n_2 + 1)S_2^2}{\sigma_2^2}}{n_1 + n_2} \right\}^{-1/2} \qquad (1.2)$$

is distributed as $t_{(n_1 + n_2)df}$.

Define $u = S_1^2/S_2^2$, $N = (n_1 + 1)/(n_2 + 1)$ and with $\theta = \sigma_1^2/\sigma_2^2$, we can rewrite (1.2) as

$$t = \frac{(d - \mu_1 + \mu_2)\ (n_1 + n_2)^{1/2} \left\{\frac{\sigma_1^2}{n_1 + 1} + \frac{\sigma_2^2}{n_2 + 1}\right\}^{-1/2}}{\left\{\frac{\frac{\sigma_2^2}{n_2 + 1}}{\frac{\sigma_2^2}{n_2 + 1}}\right\}^{1/2} \left\{\frac{n_1(n_1 + 1)S_1^2}{\sigma_1^2} + \frac{n_2(n_2 + 1)S_2^2}{\sigma_2^2}\right\}^{1/2} \left\{\frac{\frac{n_2(n_2 + 1)S_2^2}{\sigma_2^2}}{\frac{n_2(n_2 + 1)S_2^2}{\sigma_2^2}}\right\}^{1/2}}$$

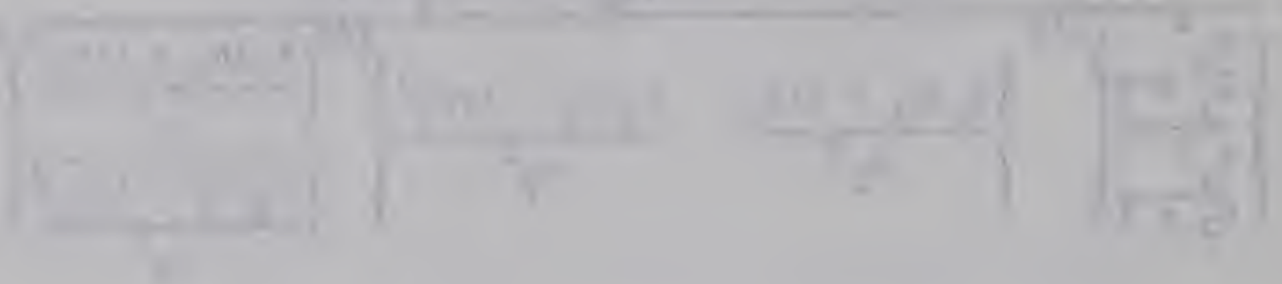
$$= \frac{(d - \mu_1 + \mu_2)\ (n_1 + n_2)^{1/2}}{\left\{\frac{\sigma_2^2}{n_2 + 1}\right\}^{1/2} \left\{1 + \frac{\theta}{N}\right\}^{1/2} \left\{\frac{n_2(n_2 + 1)S_2^2}{\sigma_2^2}\right\}^{1/2} \left\{1 + \frac{n_1}{n_2} \cdot \frac{Nu}{\theta}\right\}^{1/2}}$$

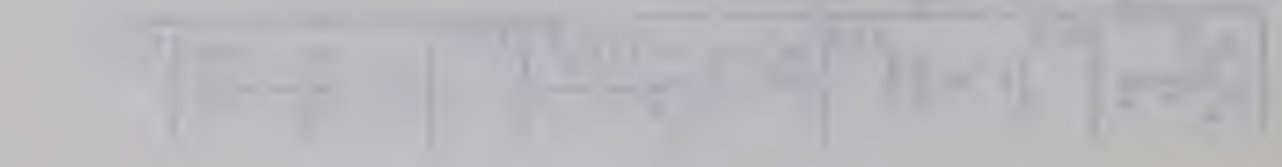
$$= \frac{(d - \mu_1 + \mu_2)\ (n_1 + n_2)^{1/2}}{n^{1/2} S_2 (1 + \frac{\theta}{N})^{1/2}\ (1 + \frac{n_1}{n_2} \cdot \frac{Nu}{\theta})^{1/2}} \quad .$$

Hence given n_1, n_2, S_1, S_2 and θ , the t statistic can be used to test the difference between means.

For Case (ii), σ_1^2 and σ_2^2 are not known or not in a known ratio. For large samples the t statistic above is asymptotically a standard normal deviate. However, the small sample solution of t involves the unknown parameter θ. Thus the problem of the test of hypothesis that the two means are equal when the true variance ratio is unknown arises.

This is the so-called Behrens-Fisher problem. In 1929 Behrens (1929) proposed an original solution which was later extended by Fisher (1935). Since then, statisticians have proposed other solutions and presented discussions on





these solutions (e.g. Welch (1938, 1947, 1956), Scheffé (1943, 1970), etc.). The controversy still remains today, forty-five years after Behrens proposed his solution, as to which solution is the best one for the Behrens-Fisher problem.

II. Objectives of Research

The objectives of the present research are three, with the second one as the major objective.

1. To give a brief survey of some of the major solutions to the Behrens-Fisher problem.

2. To examine the Behrens-Fisher test with emphasis on the method of verification of the test in Fisher (1961), and

   (a) Calculate the size of the test for selected tabular values of d.

   (b) Investigate the power of the test.

3. To attempt to reach some conclusions on the range of n_1 and n_2 in which the Behrens-Fisher test is best suited for use.

CHAPTER II. SURVEY OF LITERATURE

I. The Behrens-Fisher Solution

Behrens (1929) gave the solution as follows: Suppose we have a sample of $n_1 + 1$ observations from a normal population, yielding a sample mean $\bar{x}_1$ and sample variance $(n_1 + 1)S_1^2$, where

$$S_1^2 = \frac{\sum_{i=1}^{n_1+1} (x_i - \bar{x}_1)^2}{n_1(n_1 + 1)} .$$

Then, if μ_1 is the true mean of the population,

$$\mu_1 = \bar{x}_1 + S_1 t_1 \qquad (2.1),$$

where t_1 is distributed in the Student's distribution with n_1 degrees of freedom.

Similarly, if a sample of $n_2 + 1$ observations is taken from a second normal population, then for the second mean, we can write $\mu_2 = \bar{x}_2 + S_2 t_2$ (2.2), where t_2 is distributed in the Student's distribution with n_2 degrees of freedom independently of t_1 and $\bar{x}_2$, S_2 are similarly defined as in the first population.

Under the null hypothesis that the two population means are equal, we have, from (2.1) and (2.2),

$$d = \bar{x}_1 - \bar{x}_2 = S_2 t_2 - S_1 t_1 \qquad (2.3).$$

Since S_1, S_2 are known from the samples, the expression on the right of (2.3) has a known distribution depending only on the distribution of t_1 and t_2.

Fisher extended Behrens' solution and derived the solution using the fiducial argument.

The principle of the fiducial argument is illustrated in Fisher (1935) by applying it to the Student's t solution.

If a sample of n observations $x_1, \ldots, x_n$ is drawn from a normal population with mean μ_1 and if from the sample we calculate two statistics,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad , \quad S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n(n-1)}$$

then $t = \frac{(\bar{x} - \mu)}{S}$ is distributed as t with n degrees of freedom. Now if p is the probability that $t > t_1$, then p is also the probability that $\mu < \bar{x} - St_1$. The probability is known for all values of t_1. Hence the probability that μ is less than any assigned value is also known. In other words, the probability distribution of μ is known. The probability distribution of μ derived in this manner is termed the fiducial distribution of μ. This fiducial argument has never been fully accepted (or perhaps understood) by statisticians, largely because μ is a parameter assumed to be constant (if unknown) and not a random variable with a probability distribution.

Fisher (1935) then goes on to the problem of the difference between two means. He let $\bar{x}_1 - \bar{x}_2 = y$ and $\mu_1 - \mu_2 = \delta$. Then $\varepsilon = y - \delta = S_2 t_2 - S_1 t_1$ (2.4). Let $\tan \theta = S_1/S_2$. Dividing both sides of (2.4) by $\sqrt{S_1^2 + S_2^2}$, we get

$$\frac{\bar{x}_1 - \bar{x}_2 - \mu_1 + \mu_2}{\sqrt{s_1^2 + s_2^2}} = \frac{s_2 t_2}{\sqrt{s_1^2 + s_2^2}} - \frac{s_1 t_1}{\sqrt{s_1^2 + s_2^2}}$$

Define $$d = \frac{\bar{x}_1 - \bar{x}_2 - \mu_1 + \mu_2}{\sqrt{s_1^2 + s_2^2}}$$

Therefore $d = t_2 \cos\theta - t_1 \sin\theta$ (2.5).

Now, Fisher's fiducial approach starts with the fact that t_1 and t_2 are independent Student variates, μ_1 and μ_2 are regarded as random variables and $\bar{x}_1$, $\bar{x}_2$, s_1^2, s_2^2 are regarded as known and fixed once a pair of samples are taken. In this manner, θ is known and d becomes a linear combination of two independent Student variates and thus has a known distribution.

When n_1 and n_2 are both increased, the distribution of d tends to be normal and independent of θ. When $\theta = 0°$ or $90°$, the distribution of d is of Student's form, with n_2 or n_1 degrees of freedom respectively. In general d involves n_1, n_2 and θ. Thus to tabulate values of d for any chosen probability, we would require a table of triple entry. Since critical values of d for any α can be calculated, we can use d to test the hypothesis that $\mu_1 - \mu_2 = 0$.

A table of the significant values of d as defined in (2.5) was calculated by Sukhatme (1938). The table covers values of even degrees of freedom n_1 and n_2 at 6, 8, 12, 24 and ∞, and values of θ at $0°$, $15°$, $30°$, $45°$, $60°$, $75°$, $90°$.

Fisher (1941) gave asymptotic expansions for calculating the probabilities of d and the critical values of d for specified values of α. He also gave a table for the case when n_1 or n_2 is large.

Fisher and Healy (1956) calculated the exact values of d for small odd degrees of freedom 1, 3, 5, 7 for α = .10, .05, .02, .01. All the three tables are reprinted in Fisher and Yates (1963).

II. Welch Approximate Degrees of Freedom (APDF) Solution

Define n_1, n_2 to be sample sizes, $\lambda_i = \frac{1}{n_i}$, $f_i = n_i - 1$, $i = 1,2$. Welch (1938) considered two criteria,

$$u = \frac{\bar{x}_1 - \bar{x}_2}{\left\{ \frac{\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2 + \sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2}{(n_1 + n_2 - 2)} \times \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right\}^{1/2}}$$

$$\text{and } v = \frac{\bar{x}_1 - \bar{x}_2}{\left\{ \frac{\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2}{n_1 f_1} + \frac{\sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2}{n_2 f_2} \right\}^{1/2}}$$

$$\text{Let } S_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ji} - \bar{x}_j)^2}{f_j}, \quad j = 1,2.$$

When $\mu_1 = \mu_2$, $\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = \chi'$,

also $\dfrac{f_1 S_1^2}{\sigma_1^2} = \chi_1^2$, $\dfrac{f_2 S_2^2}{\sigma_2^2} = \chi_2^2$

where χ'^2, χ_1^2, χ_2^2 are distributed as χ^2 with 1, f_1, f_2 degrees of freedom respectively.

It is possible to write u and v in the form

$$y = \frac{\chi'}{\sqrt{a\chi_1^2 + b\chi_2^2}} = \frac{\chi'}{\sqrt{w}} \qquad (2.6)$$

where a and b are constants depending on the n'S and σ'S and w is distributed independently of χ'.

The distribution of w is then approximated by the Pearson Type III curve with

$$p(w) = \frac{w^{f/2 - 1}}{(2g)^{f/2} . e^{w/2g} . \Gamma(f/2)} \qquad (2.7)$$

where f and g are chosen so that the first two moments of the curve agree with the true moments of w. The values of f and g are found to be

$$f = \frac{(a f_1 + b f_2)^2}{a^2 f_1 + b^2 f_2}, \qquad g = \frac{a^2 f_1 + b^2 f_2}{a f_1 + b f_2} \qquad (2.8)$$

$\frac{w}{g}$ is distributed approximately as chi square with f degrees

of freedom. Hence $\chi'(w/fg)^{-1/2} = t_f$ (2.9) is distributed approximately as t with f degrees of freedom. Therefore, from (2.6) we have $y = c\, t_f$, with $c = (fg)^{-1/2} = (af_1 + bf_2)^{-1/2}$, and f is given by (2.8). Now u and v are of the form (2.9) and a and b for the two criteria can be found. Hence c and f can be found. For instance, for v, $c = 1$ and

$$f = \frac{(\lambda_1\sigma_1^2 + \lambda_2\sigma_2^2)^2}{\frac{\lambda_1^2\sigma_1^4}{f_1} + \frac{\lambda_2^2\sigma_2^4}{f_2}}$$

Since σ_i^2 are not known, Welch gave an estimate of f as

$$f = \frac{(\lambda_1 S_1^2 + \lambda_2 S_2^2)^2 - 2\left(\frac{\lambda_1^2 S_1^4}{n_1 + 1} + \frac{\lambda_2^2 S_2^4}{n_2 + 1}\right)}{\frac{\lambda_1^2 S_1^4}{n_1 + 1} + \frac{\lambda_2^2 S_2^4}{n_2 + 1}} .$$

Welch also stated that when it can be assumed that $\sigma_1 = \sigma_2$, then u can be used. But if $\sigma_1 \neq \sigma_2$, u would be biased and it is better to use v.

This solution has the obvious advantage that it only involves referencing the criterion u or v to the Student's t table which is readily available, with degrees of freedom given by f.

III. Welch-Aspin Solution

Welch (1947) developed an approximate series solution for the Behrens-Fisher problem. He was concerned with finding a quantity h, calculated from the observed variances and depending on the size of the test p, such that

$$P_r[\{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)\} < h(S_1^2, S_2^2, p)] = p \quad (2.10)$$

The solution is developed generally for k populations, with the Behrens-Fisher problem a particular case at k = 2. Let us describe briefly the solution for the case k = 2.

Again, let n_1, n_2 be the sample sizes, $\lambda_i = \frac{1}{n_i}$, $f_i = n_i - 1$, i = 1,2. Let $j(S_1^2, S_2^2, p)$ be the probability that $(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)$ is less than $h(S_1^2, S_2^2, p)$ given S_1^2 and S_2^2. Now since $\bar{x}_1 - \bar{x}_2$ is distributed independently of S_1^2 and S_2^2, we have

$$j(S_1^2,S_2^2,p) = \int_{-\infty}^{h(S_1^2,S_2^2,p)(\lambda_1\sigma_1^2 + \lambda_2\sigma_2^2)^{-1/2}} (2\pi)^{-1/2} e^{-u^2/2} \, du$$

To satisfy the condition of (2.10), we simply average $j(S_1^2,S_2^2,p)$ over the known probability distribution of S_i^2 and the result will equal p. So

$$\int_{S_i^2} \int j(S_1^2,S_2^2, p) \prod_{i=1}^{2} p(S_i^2) d S_i^2 = p \quad (2.11)$$

Welch then proceeded to develop a series expansion for $h(S_1^2,S_2^2,p)$ using the relation (2.11). The solution developed (to the order $\frac{1}{f_i^2}$) is:

$$h(S_1^2,S_2^2,p) = \xi \sqrt{\sum_{i=1}^{2} \lambda_i S_i^2} \left\{ 1 + \frac{(1+\xi)^2}{4} \frac{\left(\sum_1^2 \frac{\lambda_i^2 S_i^4}{f_i}\right)}{\left(\sum_1^2 \lambda_i S_i^2\right)^2} \right.$$

$$- \frac{(1+\xi)^2}{2} \frac{\left(\sum_1^2 \frac{\lambda_i^2 S_i^4}{f_i^2}\right)}{\left(\sum_1^2 \lambda_i S_i^2\right)^2} + \frac{(3+5\xi+\xi^4)}{3} \frac{\left(\sum_1^2 \frac{\lambda_i^3 S_i^6}{f_i^2}\right)}{\left(\sum_1^2 \lambda_i S_i^2\right)^3}$$

$$- \frac{(15+32\,\xi^2+9\,\xi^4)}{32} \frac{\left(\sum_1^2 \frac{\lambda_i^2 S_i^4}{f_i}\right)^2}{\left(\sum_1^2 \lambda_i S_i^2\right)^4} \Bigg\}$$

where ξ is the standard normal deviate such that

$$\int_{-\infty}^{\xi} (2\pi)^{-1/2} \cdot e^{-u^2/2}\, du = p \quad .$$

Aspin (1948,1949) extended the series expansion for the Welch test and tabulated values of the v statistic as a function of the observed variance ratio

$$C = \frac{\lambda_1 S_1^2}{\lambda_1 S_1^2 + \lambda_2 S_2^2} \quad .$$

Two-sided .10 and .02 critical values or one-sided .05 and .01 values are available at f_1, f_2 = 6, 8, 10, 15, 20, ∞. James, Trickett and Welch (1954,1956) calculated two-sided .05 and .01 or one-sided .025 and .005 critical values of the v statistic also in terms of C and at the same degrees of freedom. Some of these tables are reprinted in Pearson and Hartley (1954).

IV. Scheffé's Solution

An exact confidence solution was given by Scheffé (1943) with the use of the t distribution. Let $(x_1, x_2, \ldots, x_{n_1})$ and $(y, \ldots, y_{n_2})$ be random samples from normal populations. Assume $n_1 \leq n_2$. The solution is

$$\frac{\bar{x}_1 - \bar{x}_2 - \mu_1 + \mu_2}{S_d} = t_\alpha (f_{n_1})$$

where $t_\alpha (f_{n_1})$ is distributed as t with n_1 degrees of freedom, $S_d = \left(\frac{Q}{n_1(n_1 - 1)} \right)^{1/2}$,

$$Q = \sum_1^{n_1} (u_i - \bar{u})^2 \quad , \quad u_i = x_i - \left(\frac{n_1}{n_2}\right)^{1/2} \cdot y_i \quad ,$$

$$\bar{u} = \sum_{i=1}^{n_1} \frac{u_i}{n_1} \quad .$$ The solution holds for any randomly selected subset of n_1 of the n_2 values in the second sample.

Scheffé (1970) in a survey of various solutions to the problem, said that his solution is impractical and thus he does not recommend its use because the calculation of S_d involves putting in random order the elements of the larger sample $(y_1, \ldots, y_{n_2})$. The value of S_d and hence the value of the test statistic $(\bar{x}_1 - \bar{x}_2 - \mu_1 + \mu_2)/S_d$ depends very much on the result of this randomization.

There are other solutions proposed.

McCullough, Gurland and Rosenberg (1960) proposed a solution using the statistic

$$Y(r_1,r_2) = \frac{(\bar{x}_1 - \bar{x}_2)^2}{r_1\Sigma_1 + r_2\Sigma_2} \quad , \quad \text{where} \quad \Sigma_i = \sum_j (x_{ij} - \bar{x}_i)^2,$$

$$i = 1,2.$$

r_1, r_2 are constants appropriately chosen to control the size of the test and the critical value for the test is a constant equal to 1. That is, in testing the hypothesis $\mu_1 = \mu_2$, the hypothesis is rejected when $Y(r_1,r_2) > 1$.

Jeffreys (1940) developed a Bayesian solution and arrived at the same distribution as the Behrens-Fisher solution. An approximation to the d distribution by the Student's t distribution has been proposed by Patil (1964). Box and Tiao (1973) studied the approximation at one set of (n_1, n_2, θ).

At the same time that various solutions were put forward, many discussions and criticisms on these solutions arose.

On the Behrens-Fisher solution, Welch (1956) disagreed with the fact that S_1/S_2 is fixed. He was of the opinion that the right approach was to average over S_1^2 and S_2^2 as he did in his series solution.

Others (Bartlett (1936), Neyman (1941)) based their arguments on the confidence interval approach and noted that in repeated sampling from populations with a fixed true variance ratio, the Behrens-Fisher test would not reject the

null hypothesis with a frequency equal to the specified size. This, they say, shows some defect in the solution since in the confidence interval approach this requirement should be satisfied for a test of hypothesis.

In answer to this, Fisher (1939,1961) clarified the hypothesis on which his test was based. This will be discussed in detail in the next chapter. Yates (1939) also pointed out that the apparent inconsistency of the solution is due to an insufficient appreciation of the fiducial basis of the solution.

Wilks (1940) stated that it can be shown that there exists no function of $\bar{x}_1$, $\bar{x}_2$, S_1^2, S_2^2, $\mu_1-\mu_2$ independent of σ_1 and σ_2 having its probability law independent of the four population parameters. Hence it is impossible to obtain exact confidence limits for $\mu_1-\mu_2$ corresponding to a given confidence coefficient. However, no proof has been published.

Bennett and Hsu (1961) conducted a sampling study on the power of the Behrens-Fisher and the Welch-Aspin test. In the study, random normal samples x_{i_1k} ($i_1 = \ldots, n_1$, $k = 1, \ldots, 100$), and y_{j_2k} ($j_2 = 1, \ldots, n_2$, $k = 1, \ldots, 100$) were generated. For each selected pair (n_1,n_2) and assigned value of the parameters μ_1, σ_1, μ_2, σ_2, the ratio

$$v_k = \frac{\bar{x}_k - \bar{y}_k}{\sqrt{\frac{s^2_{1k}}{n_1} + \frac{s^2_{2k}}{n_2}}} \quad (k = 1, \ldots, 100)$$

was computed for the 100 samples, where $\bar{x}_k$, $\bar{y}_k$, s^2_{1k}, s^2_{2k} are the sample means and variances. The power of this test is the relative frequency of the event $\left\{v_k \leq -v_{\frac{\varepsilon}{2}} \mid \mu_1 < \mu_2\right\}$ where ε is the specified size. Assigned values of μ_1, μ_2 are $\mu_1 = 0$, $\mu_2 = 0, 0.5, 1.0, 2.0, 3.0$, and the interpretation of the significance or nonsignificance of each v_j was determined by the use of the Fisher and Yates tables and the Pearson-Hartley tables respectively for the two tests.

The results show that for smaller values of n_1 and n_2 the Behrens-Fisher test showed a smaller empirical size (power of the test at $\mu_1 = \mu_2$) than the Welch test. The power of the Welch test is greater than the Behrens-Fisher test over the whole range of $\mu_1 - \mu_2$.

Yates (1964) reviewed some of the concepts of the fiducial theory and the arguments for and against the Behrens-Fisher solution. He pointed out the importance of recognizing the proper reference set of a test of significance.

Mehta and Srinivasan (1970) and Wang (1971) also calculated the actual size of the Behrens-Fisher solution together with other solutions. The actual size of the test at various fixed true variance ratios and selected (n_1, n_2) is found to be lower than the nominal size.

Thus proceeding along different lines of reasoning, Fisher, Welch, Scheffé etc., arrived at different solutions to the same problem. It appears that the Behrens-Fisher solution is the most often discussed since Behrens (1929)

and Fisher (1935) first put it forward and Sukhatme (1938) published tables for using the test. We are drawn to this solution ourselves and will examine it in more detail in this study.

CHAPTER III. VERIFICATION OF THE BEHRENS-FISHER TEST

In reply to the criticisms to the Behrens-Fisher solution, Fisher (1939, 1961) clarified the hypothesis on which the test was based and gave methods of verification of the test based on the hypothesis. The important points in Fisher (1939,1961) are given below.

A. Fisher (1939)

I. The distribution of the d^2 statistic is expressed in terms of t^2 where t has the Student's distribution with $n_1 + n_2$ degrees of freedom. The probability with which the tabulated values of d are exceeded by the means of samples from populations having the same mean can then be calculated using this relationship with t.

Define

$$S_1^2 = \frac{\sum_{i=1}^{n_1+1} (x_{1i} - \bar{x}_1)^2}{n_1(n_1+1)}$$

$$S_2^2 = \frac{\sum_{i=1}^{n_2+1} (x_{2i} - \bar{x}_2)^2}{n_2(n_2+1)} \quad ,$$

v_1, v_2 to be the true variances of the population means, and σ_1^2, σ_2^2 the true population variances.

Suppose the pair of samples come from populations having the same mean. Therefore we have

$$\bar{x}_1 - \bar{x}_2 \sim N(0, v_1 + v_2) \text{ or}$$

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{v_1 + v_2}} \sim N(0,1) \quad .$$

Since $\frac{n_1 S_1^2}{v_1} \sim \chi^2_{n_1 df}$, $\frac{n_2 S_2^2}{v_2} \sim \chi^2_{n_2 df}$ and

$$\left(\frac{n_1 S_1^2}{v_1} + \frac{n_2 S_2^2}{v_2}\right) \sim \chi^2_{(n_1 + n_2)df}$$

$$\text{Define } t^2 = \frac{(\bar{x}_1 - \bar{x}_2)^2/(v_1 + v_2)}{\left(\frac{n_1 S_1^2}{v_1} + \frac{n_2 S_2^2}{v_2}\right) \Big/ (n_1 + n_2)} \qquad (3.1)$$

Then t has the Student's t distribution with $n_1 + n_2$ degrees of freedom. Let $w = \frac{v_1}{v_2}$ and $d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S_1^2 + S_2^2}}$ as before.

Multiplying the right hand side of (3.1) by $\frac{v_1}{v_1}$ we get

$$t^2 = \frac{d^2(S_1^2 + S_2^2)\ (n_1 + n_2)}{(1 + \frac{1}{w})\ (n_1 S_1^2 + n_2 S_2^2 w)} \qquad (3.2)$$

Expression (3.2) clearly involves w which is unknown. However, w varies according to a certain probability law. Let $F = S_1^2 V_2/(S_2^2 V_1)$. F has the Snedecor's F distribution with n_1 and n_2 degrees of freedom. Hence, the random variable $z = \frac{1}{2} \log F = \frac{1}{2} \log S_1^2 (S_2^2 w)^{-1}$ has the known z distribution with n_1 and n_2 degrees of freedom. With S_1^2/S_2^2 known

from the sample, w is distributed as $S_1^2/(S_2^2 e^{2z})$. The probability with which a particular tabulated d value will be exceeded is then the average value of the probabilities with which it will be exceeded for various possible values of w.

In short, the procedure is as follows: In order to obtain a good spread of the values of w, let p_i be a member of a series of uniformly spaced fractions, such as $\frac{1}{32}$, $\frac{3}{32}$, ..., $\frac{31}{32}$. Carry out steps a-d for each p_i.

a. Calculate w.

b. Calculate $(t_{p_i})/d$ using (3.2).

c. Calculate t_{p_i} using the $d_{\alpha/2}$ value from Sukhatme's table.

d. Calculate $P(t > t_{p_i}) = A$. This is the probability with which t_{p_i} will be exceeded.

e. Take the average of the values of A calculated from (d). The average of these values should be $\alpha/2$.

Fisher illustrated this procedure with an example. The case he considered was $n_1 = n_2 = 6$, $S_1 = S_2$, $\alpha/2 = .025$. Values of $p_i = \frac{1}{32}, \frac{3}{32}, \ldots$, etc. were used. The probability with which the tabulated d value is exceeded is calculated to be .0238. It is suggested that a finer graduation would increase the contribution of the tails and yield a value closer to .025.

II. Values of d appropriate to different suppositions on the true variances are given.

Consider three cases:

(1) true variances are equal ($\sigma_1^2 = \sigma_2^2$).

That is, $w = \frac{n_2 + 1}{n_1 + 1}$. From (3.2) we have

$$d^2 = \frac{(1 + \frac{1}{w})\ (n_1 S_1^2 + n_2 S_2^2 w)\, t^2}{(n_1 + n_2)\ (S_1^2 + S_2^2)}.$$

Substituting $w = \frac{n_2 + 1}{n_1 + 1}$ into the above formula, we get

$$d^2 = \frac{(n_1 + n_2 + 2)\ (n_1(n_1 + 1)S_1^2 + n_2(n_2 + 1)S_2^2)}{(n_1 + n_2)\ (n_1 + 1)\ (n_2 + 1)\ (S_1^2 + S_2^2)}\, t^2 \qquad (3.3)$$

Note that $d^2 = t^2$ when $n_1 = n_2$.

If $n_1 = 6$, $n_2 = 8$, the .05 value of t for 14 df is 2.145.

Let S_1^2/S_2^2 equal 3, 1, $\frac{1}{3}$ successively, we get from (3.3), d = 2.033, 2.109, 2.320.

(2) w is exactly equal to the sample variance ratio. That is, $w = S_1^2/S_2^2$. From (3.2) we have

$$t^2 = \frac{d^2(n_1 + n_2)\ (S_1^2 + S_2^2)}{\left(1 + \frac{S_2^2}{S_1^2}\right)\ \left(n_1 S_1^2 + \frac{S_1^2}{S_2^2} n_2 S_2^2\right)}$$

$$= \frac{d^2(n_1 + n_2)\ (S_1^2 + S_2^2)}{\left(\frac{S_1^2 + S_2^2}{S_1^2}\right)\left(\frac{S_1^2 S_2^2 (n_1 + n_2)}{S_2^2}\right)}$$

$$= d^2.$$

(3) w differs from the sample variance ratio S_1^2/S_2^2 by sampling errors given by the z distribution. In this case the d values are those given in Sukhatme's table.

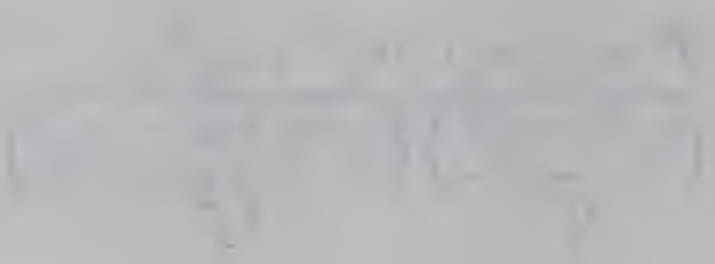
At $\alpha = .05$ the values of d appropriate to the three cases are:

S_1^2/S_2^2	3	1	1/3
Case 1	2.033	2.109	2.320
Case 2	2.145	2.145	2.145
Case 3	2.398	2.364	2.332

From these two points in the paper, it is now clear that the Behrens-Fisher test was based on the hypothesis that the true variance ratio w is not fixed nor equal to the sample variance ratio but varies according to a certain probability law. Moreover, from the second point, by assuming otherwise, one is in fact dealing with values of d other than those in Sukhatme's table.

A further note is that Fisher (1956) derived the exact sampling distribution of d as

$$\frac{d^2}{t^2} = \frac{(1 + e^{2z} \cot^2\theta)\ (n_1 + n_2\ e^{-2z})}{(n_1 + n_2)\ \mathrm{cosec}^2\ \theta}$$



which is the same as (3.2) when $S_1^2/S_2^2 e^{2z}$ and tan θ are substituted for w and S_1/S_2 in (3.2).

B. Fisher (1961)

Fisher (1961) outlined a method of verification using random sampling. He wrote that the first step in understanding a test of significance and hence in setting up a process of verification of the test is the recognition of the appropriate reference set of the test. (Reference set was defined in Fisher (1959) as the population for which probability statements of the test are made). The reference set in this test is characterized by the known values n_1, n_2 and S_1/S_2. Hence to set up a sampling process to verify the tabulated d values of the Behrens-Fisher test, the first step is to be able to obtain random samples of sizes $n_1 + 1$, $n_2 + 1$ respectively and having the correct value of S_1/S_2.

The method of verification has the same approach as the one in Fisher (1939) in the sense that the true variance ratio is also not assumed fixed based on the hypothesis by which the test was furnished. We outline the method in the following steps:

(1) Let p_i be a series of fractions, such as

$$p_i = \frac{(2i - 1)}{20,000} \quad , \quad (i = 1,2,\ldots,10000)$$

(2) If σ_1^2, σ_2^2 are as defined on p.19, the distribution

of $z = \frac{1}{2} \log \frac{(n_1 + 1)S_1^2\sigma_2^2}{(n_2 + 1)S_2^2\sigma_1^2}$ is known in terms of n_1 and n_2. Let z_p stand for the value such that $P(z < z_p) = p_i$. Calculate z_{p_i} for each p_i. Therefore for each i, i = 1, 10000, the true variance ratio

$$\frac{\sigma_2^2}{\sigma_1^2} = \frac{(n_2 + 1)S_2^2}{(n_1 + 1)S_1^2} \exp(2z_{p_i})$$

is known.

(3) Take samples of sizes $n_1 + 1$, $n_2 + 1$ respectively from two normal populations having equal means and variances in the ratio calculated on step 2.

Calculate S_1/S_2 of the samples taken. Reject the pair of samples if the ratio S_1/S_2 does not agree within a specified tolerance, with the given value of S_1/S_2. For each value of i, the first that satisfies this condition is taken as a representative sample of the reference set.

(4) To verify the tabulated d_p value corresponding to the given n_1, n_2, S_1/S_2, find the number of samples falling in the three regions:

$$\bar{x}_1 - \bar{x}_2 < - d_p \sqrt{S_1^2 + S_2^2}$$

$$- d_p \sqrt{S_1^2 + S_2^2} \leq (\bar{x}_1 - \bar{x}_2) \leq d_p \sqrt{S_1^2 + S_2^2}$$

$$d_p \sqrt{S_1^2 + S_2^2} < (\bar{x}_1 - \bar{x}_2)$$

where $\bar{x}_1, \bar{x}_2, S_1, S_2$ are calculated from the pair of samples taken. The expected numbers in these three regions are

250, 9500, 250 for the $\alpha = .05$ point of d and 50, 9900, 50 for the .01 point of d.

C. Conclusion

The Behrens-Fisher solution, arising from the fiducial argument, has the following suppositions:

1. The observed means $\bar{x}_1$, $\bar{x}_2$ and observed variance ratio S_1/S_2 are regarded as known quantities once a pair of samples are taken and leads to the solution.
2. The unknown true variance ratio w is distributed as $S_1^2/S_2^2 e^{2z}$ where z has a known distribution.

Fisher (1939,1961) emphasized that a proper verification of his test consist of an understanding of the hypothesis underlying the test. From the literature that we have covered, it is clear that most of the criticisms of the test are that the actual size of the test is lower than the nominal size. However, no mention was ever made of the clarification and verification procedure of the test in the two papers discussed above.

In this present study, we do not intend to enter into an involved discussion of the concept of fiducial inference. The merit of a test of significance is usually measured by the closeness of its actual size to the nominal size and its power performance. Since these are what most of the critics of the test are concerned with, we decided to conduct a verification of the test through the calculation of the

actual size and its power.

A proper verification of the Behrens-Fisher test based on its suppositions is contained in Fisher (1939, 1961). In the present research, we will be mainly concerned with a sampling study on the actual size of the test based on Fisher (1961).

A description of the implementation of the verification and a tabulation and discussion of the results on actual size are presented in Sections A and C in the next chapter. Actual size calculated using (3.2) is also given in Section C. Section D contains calculation of the power of the test.

CHAPTER IV. IMPLEMENTATION OF FISHER'S PROPOSED VERIFICATION OF THE BEHRENS-FISHER TEST

A. Description of Algorithm

We describe below an algorithm to implement the verification of the Behrens-Fisher test proposed by Fisher (1961).

Step 1. Calculate z_{p_i} for each p_i, where p_i is a series of fractions ($p_i = \frac{2i - 1}{2N}$, $i = 1, 2, \ldots, N$) and z_{p_i} has the known z distribution with n_1, n_2 degrees of freedom.

Calculation of z_{p_i} involves the standard normal deviate corresponding to the same p_i. Two approximations are used in order to calculate z_{p_i} for each p_i.

1. An approximation to calculate a standard normal deviate y_p satisfying

$$p_i = \frac{1}{\sqrt{2\pi}} \int_{y_p}^{\infty} e^{-t^2/2}\, dt = Q(y_p) \ .$$

That is, calculate y_p such that $Q(y_p)$ is the upper tail of a standard normal cumulative distribution function. We use the inverse normal approximation due to Hastings (1955). This method gives a $\hat{y}_p$ which approximates y_p satisfying

$$p = \frac{1}{\sqrt{2\pi}} \int_{y_p}^{\infty} e^{-\frac{t^2}{2}}\, dt \quad \text{for } 0 < p \leq 0.5 \tag{4.1}$$

The value y_p which satisfies equation (4.1), in terms of the approximation and associated error, is

$$y_p = t - \frac{c_0 + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2 + d_3 t^3} + \varepsilon$$

where $|\varepsilon| < 4.5 \times 10^{-4}$.

$c_0 = 2.515517$ $\quad$ $d_1 = 1.432788$

$c_1 = 0.802853$ $\quad$ $d_2 = 0.189269$

$c_2 = 0.010328$ $\quad$ $d_3 = 0.001308$

and $t = (-2 \ln p)^{1/2}$

For $p > 0.5$, use $1 - p$ and the y_p we want is $-y_p$.

2. An approximation to calculate z_p such that $P(z < z_p) = p_i$.

The approximation is due to Cornish and Fisher (1937). The value of z_p corresponding to a probability α, that is $P(z > z_p) = \alpha$, is expressed in terms of a standard normal deviate ξ corresponding to the same probability.

We have

$$z_p = \xi(\sigma/2)^{1/2} - \tfrac{1}{6}\delta(\xi^2 + 2) + (\sigma/2)^{1/2}\left\{(\xi^3 + 3\xi)\sigma/24 + (\xi^3 + 11\xi)\delta^2/\sigma 72\right\} - (\xi^4 + 9\xi^2 + 8)\,\delta\sigma/120 + (3\xi^4 + 7\xi^2 - 16)\,\xi^3/3240\sigma + (\sigma/2)^{1/2}\left\{(\xi^5 + 20\xi^3 + 15\xi)\,(\sigma^2/1920) + ((\xi^5 + 44\xi^3 + 183\xi)\,\delta^4/2880) + (9\xi^5 - 284\xi^3 - 1513\xi)\,\delta^4/155520\sigma^2\right\}$$

where

$$\sigma = n_1^{-1} + n_2^{-1}, \quad \delta = n_1^{-1} - n_2^{-1}$$

Therefore, to calculate z_{p_i} for each p_i such that $P(z < z_{p_i}) = p_i$, we do the following

(1) Compute p_i for i = 1, ..., N, where N is the number of pairs of samples generated.

(2) Let $q_i = 1 - p_i$ so that $q_i = P(z > z_{p_i})$.

(3) Calculate y_p such that $Q(y_p) = q_i$, using the first approximation as described previously.

(4) Calculate z_{p_i} such that $P(z \geq z_{p_i}) = q_i$ using the second approximation with $\xi = y_p$.

Hence we have z_{p_i} for i = 1, ..., N such that $P(z \geq z_{p_i}) = q_i$ or $P(z < z_{p_i}) = 1 - q_i = p_i$.

Step 2. Calculate the true variance ratio with n_1, n_2, S_1^2/S_2^2 known for each i .

True variance ratio $K = \dfrac{\sigma_2^2}{\sigma_1^2} = \dfrac{(n_2 + 1)S_2^2}{(n_1 + 1)S_1^2} e^{2z_{p_i}}$.

Step 3. Generate random samples from two normal populations with equal means and with variances in the ratio K just calculated in Step 2. There are two ways of generating random samples from the two specified distributions. Assume first that $\mu_1 = \mu_2 = 0$, $\sigma_1^2 = 1$, and so $\sigma_2^2 = K\sigma_1^2 = K$ and $S_1^2/S_2^2 = C$ is known.

a. The first way is to generate samples x_{1i}'s of size $n_1 + 1$ from a normal population with mean zero and unit variance. Next generate samples x_{2i}'s of size $n_2 + 1$ from a normal population with zero mean and variance σ_2^2 . Compute sample means $\bar{x}_1$ and $\bar{x}_2$ and sample variances S_1^2 and S_2^2.

b. The second way is a much faster way and is based on the independence of the numerator and the denominator of the d statistic. It is adapted from the master's thesis by West (1967). This method is used in this research. The d statistic is defined as

$$d = \frac{\bar{x}_1 - \bar{x}_2 - \mu_1 + \mu_2}{\sqrt{S_1^2 + S_2^2}}$$

(i) Consider the numerator $\bar{x}_1 - \bar{x}_2 - \mu_1 + \mu_2$.

From Chapter 1, we have $\bar{x}_1 - \bar{x}_2 \sim N(\mu_1 - \mu_2,$

$$\frac{\sigma_1^2}{n_1 + 1} + \frac{\sigma_2^2}{n_2 + 1}).$$

Again set $\mu_1 = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = K\sigma_1^2 = K$,

then $\bar{x}_1 - \bar{x}_2 \sim N(-\mu_2, \frac{1}{n_1 + 1} + \frac{K}{n_2 + 1})$.

Hence, for the numerator, we generate a normal random deviate with specified mean $-\mu_2$ and variance

$$\frac{1}{n_1 + 1} + \frac{K}{n_2 + 1}.$$

(ii) Consider the denominator $\sqrt{S_1^2 + S_2^2}$

$$\text{with } S_1^2 = \frac{\sum_{i=1}^{n_1 + 1} (x_{1i} - \bar{x}_1)^2}{n_1(n_1 + 1)}$$

where each $x_{1i} \sim N(\mu_1, \sigma_1^2)$, then it is known that

$$\frac{n_1(n_1 + 1)S_1^2}{\sigma_1^2}$$

has a χ^2 distribution with n_1 degrees of freedom. Thus to generate random numbers from the distributions of S_1^2 and S_2^2, we can generate samples from χ^2 distributions with n_1 and n_2 degrees of freedom respectively, since generating from the χ^2 distribution is much easier. Then since

$\frac{n_1(n_1 + 1)S_1^2}{\sigma_1^2}$ is $\chi^2_{n_1 df}$ and $\frac{n_2(n_2 + 1)S_2^2}{\sigma_2^2}$ is $\chi^2_{n_2 df}$,

then if the two chi square numbers generated are y_{1i} and y_{2i}, by the transformations $1/n_1(n_1 + 1)$, and $K/n_2(n_2 + 1)$,

$$\frac{y_{1i}}{n_1(n_1 + 1)} \quad \text{and} \quad \frac{y_{2i}\ K}{n_2(n_2 + 1)}$$

are then random observations from distributions of S_1^2 and S_2^2 respectively.

Step 4. Calculate S_1^2/S_2^2. Accept the pair of samples if

$$\frac{\left| \frac{S_1^2}{S_2^2} - C \right|}{C} \leq \varepsilon,$$

where ε is the specified tolerance limit, otherwise generate another observed value for S_1^2/S_2^2.

Step 5. If d_p is the tabular value to be verified, count the number of observations falling in the regions

$$\bar{x}_1 - \bar{x}_2 < - d_p \sqrt{S_1^2 + S_2^2}$$

$$- d_p \sqrt{S_1^2 + S_2^2} \leq \bar{x}_1 - \bar{x}_2 \leq d_p \sqrt{S_1^2 + S_2^2}$$

$$d_p \sqrt{S_1^2 + S_2^2} < \bar{x}_1 - \bar{x}_2 .$$

The total number of observations falling in the first and third regions divided by N gives the actual size of the two tailed test.

B. Pseudo-Random Number Generation

Generation of random numbers from certain distributions, for example, a normal distribution with specified mean and variance, or a chi square distribution with n degrees of freedom, requires first of all generation of random numbers from a U(0,1) distribution. Random numbers from various distributions are then obtained by doing some transformation to the U(0,1) random numbers. Hence, a good random number generator to generate U(0,1) numbers is a basic requirement for most sampling studies.

In this study, we need to generate random numbers from normal and chi square distributions. This involves generation of random numbers from a U(0,1) distribution. Moreover, generation of a single chi square random number

often requires more than one U(0,1) number, the number required being dependent on the degrees of freedom of the chi square distribution. Also, in our algorithm, very large numbers of random numbers are needed because for each of the 5000 pairs of samples taken, we have to keep on generating samples, rejecting those whose S_1/S_2 ratio lies outside the tolerance limit and accepting the first pair of samples whose S_1/S_2 ratio is within the tolerance limit. Thus we need a good and fast uniform random number generator to start with.

By a fast pseudo-random number generator, we mean the the generating time per number should be reasonably short. By a good pseudo-random uniform number generator we mean one that:

(1) generates numbers whose distribution approximates closely the U(0,1) distribution.

(2) satisfies various statistical criteria of randomness.

(3) generates a sequence of numbers long enough before the numbers repeat themselves.

To satisfy (1), the chi square goodness of fit test or the Kolmogorov-Smirnov test is usually performed. For (2), one performs various statistical tests on the random numbers. A detailed discussion on the various tests of randomness is found in Jansson (1966). (3) involves careful choice of constants in the mathematical equations used to generate the pseudo-random numbers.

A number of existing generators were studied. Among these, some are designed for computers other than the IBM/360 (see Downham and Roberts (1967)). Some are very fast but results on the statistical tests on the generator are not available. (See for example, Seraphin (1969)). After careful consideration, we decided that the pseudo-random number generator that is more suitable for our study is the one by Lewis, Goodman and Miller (1969).

This generator employs the frequently used multiplicative congruential method. In this method, the set of numbers x_i is generated by the equation

$$x_{i+1} = Ax_i \pmod{p},$$

and the sequence $\left\{\frac{x_i}{p}\right\}$ is then taken to be the uniform random number sequence. For this particular generator, $A = 7^5$ and $p = 2^{31} - 1$.

This generator was chosen because it was extensively tested and the tests were done on the IBM 360/67 computer. The various tests done were described and test results presented in the same paper show that the generator is remarkably good. Moreover, an Assembler language program of the generator is available and may be called conveniently in a Fortran program by the statement CALL RANDOM (INT, REAL). Here INT is any full word integer variable which should be given an initial value and REAL, the random number generated, is a full word real variable (single precision). Since it is written in Assembler language, this generator is much

faster than other generators written in Fortran or other high level languages. We have an upper bound of approximately 31.2 μsec to call a random number in the 360/67. It is thus for the above reasons that we chose this generator.

Normal Random Number Generation

As discussed in Section A, we need to generate a normal random deviate for the numerator of the d statistic. The method used is the widely used Box and Muller method found in Box and Muller (1958).

Random normal deviates are obtained by first generating two uniform random numbers U_1, U_2 on the interval (0,1). Then calculate

$$x_1 = (-2 \ln U_1)^{1/2} \cos 2\pi U_2$$

$$x_2 = (-2 \ln U_1)^{1/2} \sin 2\pi U_2$$

(x_1, x_2) will be a pair of independent random variables from the same normal distribution with mean zero and unit variance. This method is convenient because x_1, x_2 are easy to calculate. Also, two random normal deviates are obtained with two U(0,1) numbers. It is exact if the uniform random numbers are accurate.

Chi square Random Number Generation

Chi square random numbers are needed to calculate the denominator of the d statistic. We use two methods to

generate the random numbers.

(1) For any degrees of freedom n:

It is easy to show that, if $U \sim U(0,1)$, then

$$Y = -2\ln U \sim \chi_2^2 .$$

Hence, observations from a χ^2 distribution with 2k degrees of freedom can be generated by adding together the k terms

$$\sum_{i=1}^{k} (-2\ln U_i).$$

For χ^2 with 2k + 1 degrees of freedom we generate k U(0,1) numbers, get the sum

$$\sum_{1}^{k} (-2\ln U_i)$$

and add the square of a normal deviate generated by the Box and Muller method. This method has the advantage that only k uniform numbers need to be generated for a χ^2 deviate with 2k degrees of freedom.

(2) For large degrees of freedom:

For chi square deviates with large degrees of freedom, the enormous amount of U(0,1) numbers that need to be generated, and hence the length of time required for each run prompted us to look for a shorter method of generating chi square random numbers.

We use the approximation by Wilson and Hilferty (1931). The approximation is: for large n, $(\chi^2/n)^{1/3}$

is normally distributed about mean $1 - \frac{2}{9n}$ and variance $\frac{2}{9n}$. Hence, to generate a chi square number with n degrees of freedom, we can first generate x_i, where $x_i \sim N(0,1)$. Then the chi squared random number is

$$\chi^2 = n\left(1 - \frac{2}{9n} + x_i \sqrt{\frac{2}{9n}}\right)^3 \qquad (4.2)$$

This method has the obvious advantage that only one N(0,1) number needs to be generated for each chi square number with n degrees of freedom.

In Wilson and Hilferty (1931), values of χ^2 are calculated for n = 1, 2, 3, 10, 30 at p = 0.80, 0.50, 0.20, 0.05, 0.01 using (4.2). Comparison with the corresponding tabular χ^2 values shows that the approximation is quite good.

Checks on the Generation of U(0,1) Random Numbers

We did just a few tests on the U(0,1) random number generator we have chosen since it was quite well tested in Lewis, Goodman and Miller (1969) and also because a thorough test requires a lot of computing time. Nevertheless, our test results indicate that the random number generator is satisfactory.

To test the random number generator, 160,000 numbers were generated. These numbers were divided into four subsamples, each of size 40,000. The tests were then performed for each subsample. The test results are summarized in Tables 4.1 and 4.2.

Table 4.1. Tests on the Uniform Random Numbers

Trial	Mean	Second Moment	Third Moment	Fourth Moment	χ^2_{260}	k-s statistic D_n^*
1	.5012	.3352	.2518	.2018	270.55	.0058
2	.4993	.3335	.2506	.2009	208.44	.0033
3	.4976	.3313	.2483	.1987	215.09	.0048
4	.4999	.3333	.2498	.1997	302.20	.0056
Expected value	.5000	.3333	.2500	.2000	298.61(α = .05) 315.99(α = .01)	.0068(α = .05) .0082(α = .01)

Table 4.2. Serial Correlation at Lags 1 to 6

Trial \ Lag	1	2	3	4	5	6
1	.0040	-.0064	-.0056	-.0082	-.0045	.0015
2	-.0065	-.0053	-.0036	.0036	.0035	.0040
3	-.0025	-.0043	-.0056	.0041	.0003	.0006
4	.0081	-.0009	-.0056	-.0018	.0079	-.0090
Expected value	0.0	0.0	0.0	0.0	0.0	0.0

Table 4.1 shows that the sample moments are close to the expected values. Two goodness-of-fit tests were performed: the χ^2 goodness-of-fit test and the Kolmogorov-Smirnov test. For the χ^2 test, the number of classes into which the numbers were grouped is determined by the Mann and Wald criterion (Mann and Wald (1942)). For N = 40,000, the number of classes was determined to be 261. Only one chi square statistic computed is significant at $\alpha = .05$. The Kolmogorov-Smirnov test is based on the maximum difference between an empirical and a specified hypothetical cumulative distribution. The test statistic is

$$D_n^* = \max|S_i(x) - F_i(x)|,$$

where $S_i(x)$ = observed cumulative frequency for class i,

$F_i(x)$ = expected cumulative frequency for class i.

D_n^* is then compared with the critical value at specified N, the total number of observations. From Massey (1951),

for N > 35, the critical values are

$$d_{.05} = \frac{1.36}{\sqrt{N}}, \quad d_{.01} = \frac{1.63}{\sqrt{N}} .$$

Thus, in our case N = 40,000, $d_{.05} = .0068$, $d_{.01} = .0082$. None of the D_n^* computed are significant at .05 level. Hence, we conclude that the generated uniform numbers have a uniform distribution over the (0,1) interval.

The serial correlation coefficients at lags 1-6 in Table 4.2 approach 0. Standardized values of the serial correlation at lag one which are approximately N(0,1) (Anderson (1942)) were calculated and none are significant at .05 level. The runs up and down and runs above and below the mean test were also performed and the observed number of runs agree closely with the expected numbers.

Thus our results also show that the generator by Lewis, Goodman and Miller (1969) is satisfactory.

Checks on the Generation of Chi Square Random Numbers

We first did a check on the range of degrees of freedom in which the second method of chi square number generation is accurate. Values of χ^2 for degrees of freedom n = 1, 3, 5, 6, 7, 8, 12, 24, 30, 60 at probability p = .99, .95, .70 .20, .05, .01, .001 are calculated using (4.2). The calculated values are compared with the theoretical values from the χ^2 tables. It is found that the approximation becomes closer and closer as n increases. For $n \geq 8$, the maximum

relative deviation from the theoretical value is 2.976% at $p = .99$ and $n = 8$. All the other deviations are less than 1%. For $n < 8$, there are some larger deviations at $n = 1, 3, 5$. Hence, the second method should only be used for $n \geq 8$ to ensure more accuracy.

Next we test how close the simulated distributions using both methods are to the theoretical chi square distribution. The χ^2 goodness-of-fit and the Kolmogorov-Smirnov tests were performed. The sample means ($\bar{x}$) and sample variances (S^2) were also calculated and compared with the expected means (μ) and variances (σ^2).

Four trials each consisting of 5000 chi square numbers were performed for each selected degree of freedom. The class width was arbitrarily fixed at 0.5 for all n. In the χ^2 goodness-of-fit test some classes were grouped to make the expected frequencies in each class ≥ 5. Calculation of theoretical χ^2 probabilities is based on the recurrence formula in Abramowitz and Stegun (1964, p. 941). The recurrence relation is

$$P(\chi^2_f > x) = P(\chi^2_{f-2} > x) + \frac{(\frac{1}{2}\chi^2)^{\frac{f}{2}-1} e^{-\chi^2/2}}{\Gamma(\frac{f}{2})}$$

Some of the results are contained in the following table and are representative of all the runs performed.

Table 4.3. Tests on the Chi square Random Numbers

(a). Method 1.

Degrees of Freedom	Trial	$\bar{x}$	S^2	χ^2	Degrees of Freedom for χ^2 Test	D_n^*
n = 5	1	4.911	9.856	27.34	34	0.0100
(μ = 5	2	4.977	9.458	29.97	35	0.0065
σ^2 = 10)	3	4.931	9.840	29.91	32	0.0158
	4	5.016	10.137	27.89	30	0.0084
n = 6	1	5.953	12.070	28.84	32	0.0169
(μ = 6	2	5.951	11.593	39.01	34	0.0102
σ^2 = 12)	3	5.977	11.724	42.63	30	0.0086
	4	5.928	11.786	40.31	31	0.0092
n = 7	1	6.985	13.824	28.46	30	0.0052
(μ = 7	2	6.956	13.098	43.31	34	0.0077
σ^2 = 14)	3	7.042	13.567	40.06	35	0.0117
	4	6.942	14.035	29.32	30	0.0143
n = 8	1	7.941	15.095	57.44	50	0.0097
(μ = 8	2	7.935	15.905	60.52	51	0.0149
σ^2 = 16)	3	7.947	15.794	32.06	44	0.0123
	4	7.997	15.441	65.45	49	0.0077

(b) Method 2.

Table 4.3. (continued)

Degrees of Freedom	Trial	$\bar{x}$	s^2	χ^2	Degrees of Freedom for χ^2 Test	D_n^*
n = 8	1	7.990	16.008	50.55	47	.0066
(μ = 8	2	7.983	15.829	68.31$^+$	48	.0118
σ^2 = 16)	3	8.058	16.244	49.94	45	.0110
	4	7.887	15.755	49.64	47	.0198$^+$
n = 12	1	12.072	24.342	49.42	46	.0096
(μ = 12	2	11.860	23.670	52.76	49	.0196$^+$
σ^2 = 24)	3	11.980	23.690	56.44	50	.0113
	4	11.987	24.092	32.40	48	.0053
n = 24	1	23.970	47.520	101.07	95	.0112
(μ = 24	2	24.087	48.914	92.70	89	.0109
σ^2 = 48)	3	23.966	48.720	98.15	92	.0069
	4	23.940	46.429	88.98	96	.0096
n = 30	1	29.967	59.549	89.18	95	.0106
(μ = 30	2	30.097	61.346	100.58	90	.0109
σ^2 = 60)	3	29.960	61.157	88.04	92	.0072
	4	29.766	59.950	91.96	94	.0208$^+$

$^+$ Significant at .05 level.

The sample means and variances are close to the expected values $\mu = n$ and $\sigma^2 = 2n$. On the average, the mean relative deviation for the two methods is .0060 for $\bar{x}$ and .0187 for S^2.

On the goodness-of-fit tests, none of the χ^2 or Kolmogorov-Smirnov statistic calculated in Method 1 are significant. The critical values for the Kolmogorov-Smirnov test are $d_{.05} = .0192$ and $d_{.01} = .0231$. In Method 2, only a few are significant at .05 level and none at .01 level. Hence we conclude that the random numbers generated using these two methods have approximately a χ^2 distribution and that the first method gives a better approximation to the expected distribution than the second method. We should thus use Method 1 to generate chi square numbers for $n \leq 8$ and use Method 2 for $n > 8$.

C. Results and Discussion on the Size of the Test

Results

The calculation of the empirical size of the Behrens-Fisher test was carried out as described in Section A. Tabular values d_p to be verified are taken from Tables VI and VI_1 in Fisher and Yates (1963). Calculations were done for degrees of freedom n_1, n_2 ranging from 3 to 24, and for θ from 15° to 75°. Parameters (n_1, n_2, θ) were chosen according to availability of the corresponding d_p value in the two tables. The sample size and tolerance limit ε were

5000 pairs of samples and $\varepsilon = 0.10$. These choices are discussed in more detail later in this section.

As described in Section A, the actual size is calculated by adding up the number of observations in the two critical regions and dividing the sum by the number of pairs of samples taken. An example of the results on the number of actual observations in each region is given in Table 4.4. Two trials were performed for each set of (n_1, n_2, θ) and the average value of the actual size was taken. These average values are presented in Table 4.5. In the table $\tan\theta$ is, as defined previously, the experimental sample variance ratio S_1/S_2. In addition, the empirical distribution of d, values of which are calculated from each pair of samples, is obtained for some sets of (n_1, n_2, θ) and the curves of the distribution are drawn. Some typical results are shown in Figures 4.1 - 4.6.

Table 4.4. The Actual Number of Observations in the Three Regions in the Case $n_1 = n_2 = 3$, $\theta = 45°$.

Nominal Size α	No. of Obs. in the left critical region	No. of Obs. in the acceptance region	No. of Obs. in the right critical region	Actual Size
.10	256	4490	254	.1020
.05	118	4763	119	.0474
.02	42	4911	47	.0178
.01	20	4958	22	.0084

Table 4.5. Actual Size of the Behrens-Fisher Test

(a) $n_1 = n_2 = 3$

	Nominal Size α			
θ	.10	.05	.02	.01
15°	.1010	.0464	.0172	.0090
30°	.1011	.0480	.0189	.0075
45°	.1020	.0474	.0178	.0084

(b) $n_1 = n_2 = 5$

	Nominal Size α			
θ	.10	.05	.02	.01
15°	.1010	.0480	.0196	.0104
30°	.1032	.0476	.0194	.0098
45°	.1004	.0524	.0186	.0096

(c) $n_1 = 5$ $n_2 = 7$

	Nominal Size α			
θ	.10	.05	.02	.01
15°	.1070	.0505	.0238	.0110
30°	.0984	.0510	.0210	.0092
45°	.1100	.0525	.0210	.0092
60°	.1030	.0482	.0172	.0084
75°	.1104	.0545	.0213	.0111

cont'd.

Table 4.5 (cont'd.)

(d) $n_1 = n_2 = 6$

	Nominal Size α	
θ	.05	.01
15°	.0487	.0088
30°	.0486	.0096
45°	.0480	.0074

(e) $n_1 = n_2 = 8$

	Nominal Size α	
θ	.05	.01
15°	.0488	.0110
30°	.0498	.0118
45°	.0498	.0099

(f) $n_1 = 6 \quad n_2 = 12$

	Nominal Size α	
θ	.05	.01
15°	.0500	.0103
30°	.0519	.0092
45°	.0508	.0100
60°	.0518	.0099
75°	.0499	.0099

cont'd.

Table 4.5 (cont'd.)

(g) $n_1 = n_2 = 12$

	Nominal Size α	
θ	.05	.01
15°	.0509	.0092
30°	.0504	.0090
45°	.0504	.0095

(h) $n_1 = 12 \quad n_2 = 24$

	Nominal Size α	
θ	.05	.01
15°	.0493	.0104
30°	.0488	.0090
45°	.0500	.0094
60°	.0491	.0097
75°	.0496	.0094

(i) $n_1 = n_2 = 24$

	Nominal Size α	
θ	.05	.01
15°	.0507	.0102
30°	.0501	.0100
45°	.0501	.0104

To summarize the results in Table 4.5 (a) to (i) and to see more clearly how close the actual sizes are to the respective nominal sizes, we calculated the deviations of each actual size from the nominal size. The maximum absolute deviation corresponding to each n_1, n_2 and nominal size are then obtained and entered in Table 4.6. Thus for $n_1 = n_2 = 3$, $\alpha = .05$ for example, from Table 4.5 (a), the deviations are -.0036, -.0020, -.0026 at $\theta = 15°$, $30°$, $45°$ respectively. A + sign indicates a positive deviation (i.e. actual size is larger than the nominal size) and a - sign indicates a negative deviation (i.e. actual size is less than the nominal size). The maximum absolute value of these three deviations is $| - .0036 |$ at $\theta = 15°$. Hence we enter -.0036 in the .05 column and 15° in the last column of the first row of Table 4.6.

The table is further divided into two subtables. Table 4.6 (a) contains the maximum deviations for small odd degrees of freedom and $\alpha = .10, .05, .02, .01$. Table 4.6 (b) contains maximum deviations for even degrees of freedom and $\alpha = .05, .01$. Table 4.7 contains the average deviations for each degree of freedom. Tables 4.6 and 4.7 follow.

Table 4.6. Table of Maximum Deviations from the Nominal Size α

(a) Small odd degrees of freedom.

Degrees of Freedom		Maximum Deviations from Nominal Size α				
n_1	n_2	.10	.05	.02	.01	θ
3	3	+.0020				45°
3	3		-.0036	-.0028		15°
3	3				-.0025	30°
5	5	+.0032				30°
5	5		±.0024			30°,45°
5	5			-.0014		45°
5	5				±.0004	15°,45°
5	7	+.0104	+0045			75°
5	7			+.0038		15°
5	7				-.0016	60°

(b) Even degrees of freedom.

Degrees of Freedom		Maximum Deviations from Nominal Size α		
n_1	n_2	.05	.01	θ
6	6	-.0020	-.0026	45°
8	8	-.0012		15°
8	8		+.0018	30°
6	12	+.0019	-.0008	30°
12	12	+.0009		15°
12	12		-.0010	30°
12	24	-.0012	-.0010	30°
24	24	+.0007		15°
24	24		+.0004	45°

Table 4.7. Average of the Deviation for Various Degrees of Freedom

Degrees of Freedom		Average Deviations from Nominal Size α	
n_1	n_2	.05	.01
3	3	.0027	.0017
5	5	.0023	.0003
5	7	.0021	.0011
6	6	.0016	.0014
8	8	.0005	.0010
6	12	.0009	.0003
12	12	.0006	.0008
12	24	.0006	.0006
24	24	.0003	.0002

Discussion

Choice of Sample Size

Trials of 500, 1000, 2000, 5000 and 10,000 pairs of samples were carried out for a few sets of parameters (n_1, n_2, θ) at the beginning to determine the best sample size to use. Sets of 500, 1000, and 2000 pairs of samples for several choices of n_1, n_2, θ were tried and the results were not stable. Some of the calculated empirical sizes deviated (absolutely) as much as 2% from the nominal size, while others were very close to it. There was increased stability of results when samples were increased to 5000 and 10,000 pairs. The results obtained were consistently close to the nominal size.

To decide between using 5000 and 10,000 pairs of samples, the results on actual size and the execution time using these two sample sizes were considered. We found that:

(a) Stability of results on actual size for both sample sizes are similar.

(b) Using 10,000 pairs takes on the average 12 minutes on each run, while using 5000 pairs takes about 5-6 minutes.

After these considerations, we decided that 5000 pairs of samples is the optimum sample size to take.

Choice of Tolerance Limit ε

The tolerance limit ε was decided by trying values of ε from .025 to .10. In the case $n_1 = n_2 = 8$ and $\theta = 30°$,

for 1000 pairs of samples, we find that the execution time is 40 seconds for $\varepsilon = 0.10$, 1.2 minutes for $\varepsilon = 0.05$, and 3 minutes for $\varepsilon = .025$. This increase in execution time with decreasing ε is expected because when ε is decreased, the tolerance limit $| S_1/S_2 - C | / C \leq \varepsilon$ is shortened. Hence there is a greater chance of samples with the S_1/S_2 ratio lying outside the limit.

A counter was set up to count the number of pairs of samples rejected before a pair of samples with the correct S_1/S_2 ratio in the tolerance limit was obtained. The average of the number of rejected samples in all the samples was then taken. The average found was 172.6 for $\varepsilon = .025$, 49.94 for $\varepsilon = .05$ and 24.08 for $\varepsilon = .10$.

With our choice of sample size to be 5000, it is too time consuming to use $\varepsilon = .025$ in all the runs. Values of $\varepsilon = .05$ and .10 were then tried using 5000 pairs of samples. The average execution time and the average number of rejected samples are 8.7 minutes and 89.17 samples respectively for $\varepsilon = .05$. For $\varepsilon = .10$ it takes 5.5 minutes or almost half the time and 55.28 rejected samples. Again, we did not find any significant difference on the actual sizes obtained. With 5000 pairs of samples, $\varepsilon = .10$ seem to be the most reasonable limit to use. Hence, in all subsequent trials, a sample size of 5000 pairs and a tolerance limit $\varepsilon = 0.10$ were used.

Use of the Chi Square Approximation

The chi square approximation or method 2 of chi

square number generation as discussed in Section B of this chapter is more accurate for larger degrees of freedom. Hence it was used to generate the chi square random numbers only for larger degrees of freedom $n_1 = n_2 = 12$, $n_1 = 12$ and $n_2 = 24$, and $n_1 = n_2 = 24$. For other degrees of freedom, Method 1 described on p.37 was used. The execution time using the chi square approximation is much shortened. It takes about 2 minutes to run 5000 pairs of samples using the approximation and 5.5 minutes without, both with the same set of n_1, n_2 and θ.

Range of n_1, n_2, θ and the d_p Values Verified

When $n_1 = n_2$, the tabular d_p value at $\theta = 15°$ and $30°$ are the same as the d_p value at $\theta = 75°$ and $60°$ respectively. Hence only the values at $\theta = 15°$, $30°$, and $45°$ are verified. When $n_1 \neq n_2$, d_p at $\theta = 60°$, $75°$ is the same as the d_p value at $30°$ and $15°$ respectively, with n_1 and n_2 interchanged. For $\theta = 0°$ or $90°$, the d statistic has the Student's t distribution with n_2 or n_1 degrees of freedom. Hence the critical values of d for these values are the t values with the corresponding degrees of freedom taken from the Student's t table. Thus, in this way the d_p values that we have verified have already covered the most part of the two tables in Fisher and Yates (1963), except for the values corresponding to n_1 or $n_2 < 3$. For these degrees of freedom, tests done to check the z approximation we used show that the calculated z_p values have a larger relative deviation

than for degrees of freedom ≥ 3. There is a large relative deviation of 27.5% for degrees of freedom $n_1 = n_2 = 1$ at the tail of the z distribution. For degrees of freedom ≥ 3, the largest relative deviation is 2.182%, while the others are less than 1%. Hence in order to obtain the correct z_p and therefore the correct σ_1^2/σ_2^2 for each p, we decided to run the trials starting from $n_1, n_2 \geq 3$.

The available percentage points of d_p for each chosen set (n_1, n_2, θ) are verified in the same program. Thus for small odd degrees of freedom the .10, .05, .02 and .01 points of d_p are verified in the same program, and for even degrees of freedom, the .05 and .01 points are verified.

On the Four Tables Presented

Table 4.4 shows a typical result on the actual number of observations falling in the critical and acceptance regions. The expected numbers in each region for the various nominal sizes (α) are:

$\alpha = .10$ - 250, 4500, 250

$\alpha = .05$ - 125, 4750, 125

$\alpha = .02$ - 50, 4900, 50

$\alpha = .01$ - 25, 4950, 25 .

Comparing the expected and actual values, we see that they are close to each other.

The results on the actual size of the test are presented in Table 4.5, 4.6 and 4.7. They can be summarized as follows:

(1) The actual size are on the whole quite close to the nominal size.

(2) There is an almost equal number of positive and negative deviations from the nominal size. This indicates that the Behrens-Fisher test does not yield significant levels much lower than α, contrary to the claims in other papers which have investigated the test.

(3) Critical values at small odd degrees of freedom on the average yield larger deviations from the nominal size than those at larger degrees of freedom as seen from Tables 4.6 and 4.7.

We therefore conclude that the Behrens-Fisher test gives actual sizes which are close to α and there is no indication that they are much lower than α. Furthermore, the performance of the test in terms of actual size is better for larger even degrees of freedom.

On the Distribution of the d Statistic

The distribution of the d statistic was studied empirically. From Figures 4.1 - 4.6, it is seen that the distribution is symmetric about zero. Also, as the number of degrees of freedom increases, the distribution becomes less steep, more spread out and approaches a bell-shaped curve. We expect the curve to approach a normal curve as

n increases since the d statistic approaches normal as n approaches infinity.

Results on Actual Size Using the Verification Procedure in Fisher (1939).

The procedure is as described in Chapter 3. The method of calculating z is the same one used in the verification based on Fisher (1961). Various values of N and hence

$$p_i \quad (p_i = \frac{2i - 1}{2N}, \; i = 1, \ldots, N)$$

were used at the beginning to determine the best N to use. The results are presented in Table 4.8. Results on actual size using

$$p_i = \frac{2i - 1}{512}, \; i = 1, \ldots, 256,$$

are summarized in Table 4.9.

Table 4.8. Probabilities That the Tabulated Values of d are Exceeded at Various N

($n_1 = n_2 = 8$, $\theta = 75°$, α = nominal size)

N	Probabilities of Exceeding d	
	$\alpha/2 = .025$	$\alpha/2 = .005$
16	.02452	.00474
32	.02465	.00474
64	.02482	.00486
128	.02491	.00492
256	.02496	.00496
512	.02500	.00500
1000	.02500	.00500

Table 4.9. Probabilities that the Tabulated Values of d are Exceeded ($p_i = (2i - 1) / 512$, $i = 1, 2, \ldots, 256$, α = nominal size) for Even Degrees of Freedom

(a) $n_1 = n_2 = 6$

θ \ $\alpha/2$	.025	.005
15°	.0249	.00495
30°	.0249	.00493
45°	.0250	.00492

(b) $n_1 = n_2 = 8$

θ \ $\alpha/2$	.025	.005
15°	.0249	.00496
30°	.0249	.00496
45°	.0249	.00495

(c) $n_1 = 12$, $n_2 = 24$

θ \ $\alpha/2$	.025	.005
15°	.0249	.00500
30°	.0250	.00499
45°	.0249	.00498
60°	.0249	.00498
75°	.0249	.00498

Table 4.10. Probabilities that the Tabulated Values of d are Exceeded ($p_i = (2i - 1)/512$, $i = 1, \ldots, 256$, α = nominal size) for Odd Degrees of Freedom

(a) $n_1 = n_2 = 3$

θ \ $\alpha/2$	.05	.025	.01	.005
15°	.0500	.0250	.00999	.00497
30°	.0501	.0251	.00999	.00494
45°	.0501	.0251	.0100	.00495

(b) $n_1 = 3$, $n_2 = 5$

θ \ $\alpha/2$	.05	.025	.01	.005
15°	.0500	.0250	.0100	.00498
30°	.0499	.0250	.00996	.00496
45°	.0499	.0249	.00994	.00493
60°	.0499	.0249	.00991	.00490
75°	.0499	.0249	.0099	.00488

(c) $n_1 = 3$, $n_2 = 7$

θ \ $\alpha/2$	.05	.025	.01	.005
15°	.0499	.0249	.00998	.00488
30°	.0499	.0249	.00994	.00493
45°	.0499	.0249	.00996	.00496
60°	.0500	.0249	.00991	.00498
75°	.0500	.0250	.0100	.00490

Discussion

In determining the best N to use, it is seen from Table 4.8 that as N increases, the actual size approaches the nominal size. N = 256 or $p_i = (2i - 1)/512$ was used because it gives close actual sizes and a further increase in N does not improve the results to a significant extent. Moreover, the time needed for execution increases rapidly as N becomes large.

From the results shown in Table 4.9 - 4.10, it is again seen that the Behrens-Fisher test yields actual sizes close to the nominal size. This verification using the relationship of d and Student's t distribution further show that the Behrens-Fisher test rejects the null hypothesis when it is true with a frequency close to the nominal size α.

A Comparison With Other Investigations of the Behrens-Fisher Test.

The more recent investigations on the test are as mentioned in Chapter 2, Bennett and Hsu (1961), Mehta and Srinivasan (1970), Wang (1971). In addition, Yates (1964) also clarified some points on fiducial probability and the Behrens-Fisher test.

In the first three papers above, the test was examined in the context of the confidence interval approach, namely, that in repeated sampling of any fixed population, the size must be equal to the frequency with which the hypothesis is

rejected. Hence, the actual size of the test was calculated for populations with various sets of arbitrarily fixed σ_1^2/σ_2^2. The actual sizes calculated were all found to be much less than the nominal size.

However, Fisher arrived at his solution using the fiducial approach, a different approach from the confidence interval approach. In the fiducial argument, the statement that "in repeated sampling of any fixed population, the level of significance must be equal to the frequency with which the hypothesis is rejected", is not a requirement for a test of significance. In particular, repeated sampling from fixed populations is foreign to the approach (see Fisher (1955)). The fiducial solution is based on the two suppositions stated at the end of Chapter 3 (p.26), and will yield actual sizes close to α when the true variance ratio is distributed in the z distribution and not fixed during the verification.

It should thus be within the context of fiducial approach that we investigate the test. From the results that we have obtained, it is clear that a proper verification based on the assumptions of the test yields sizes close to α. We thus conclude that the Behrens-Fisher test is recommendable for practical use for the following two reasons:

(1) The test rejects the hypothesis when it is in fact true with a frequency close to the nominal size α.

(2) It is convenient to use because tables of critical values of d are available for a relatively wide range of (n_1, n_2, θ), including small odd degrees of freedom.

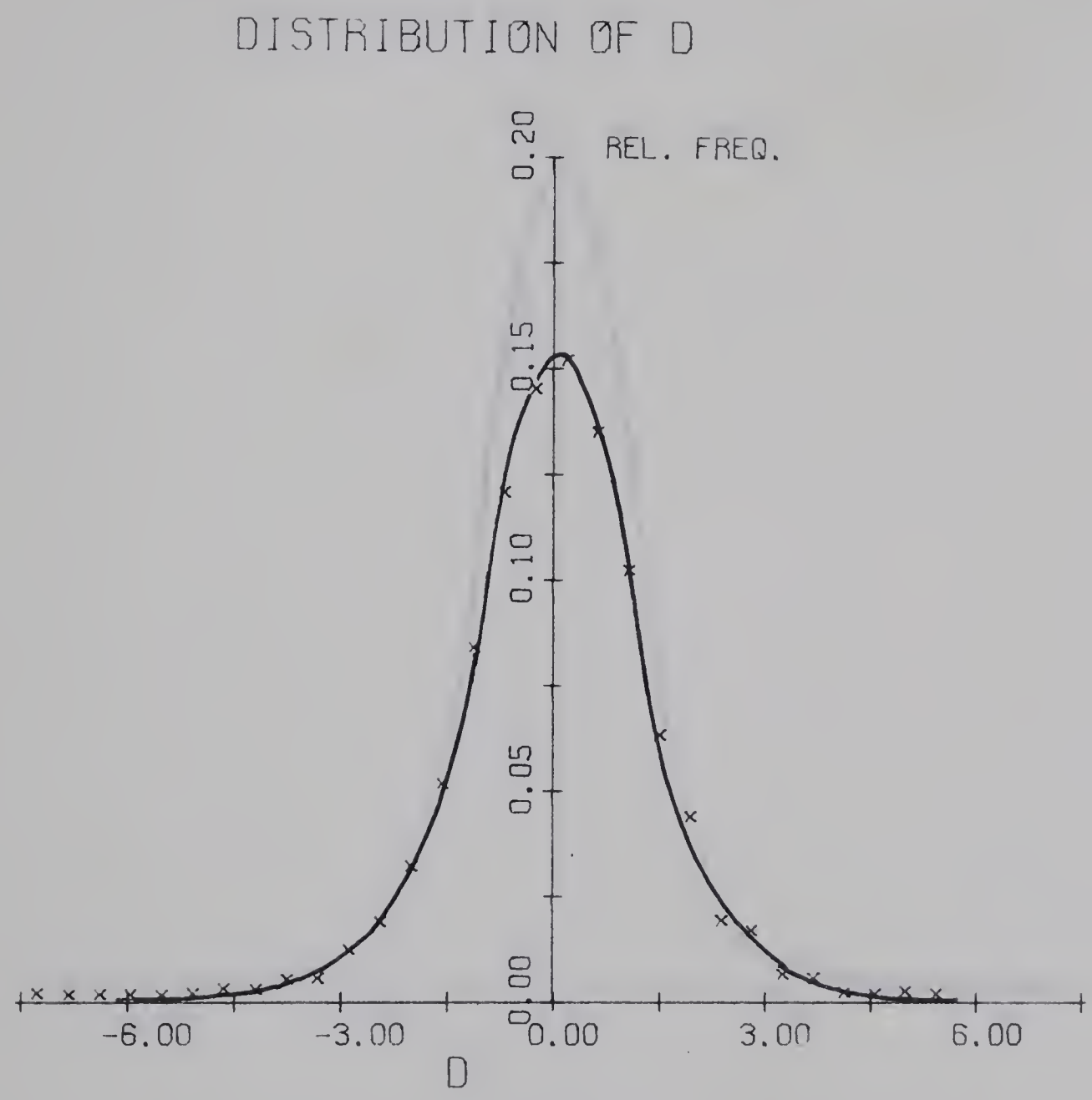


Figure 4.1. Distribution of d at $n_1 = n_2 = 5$, $\theta = 45°$.

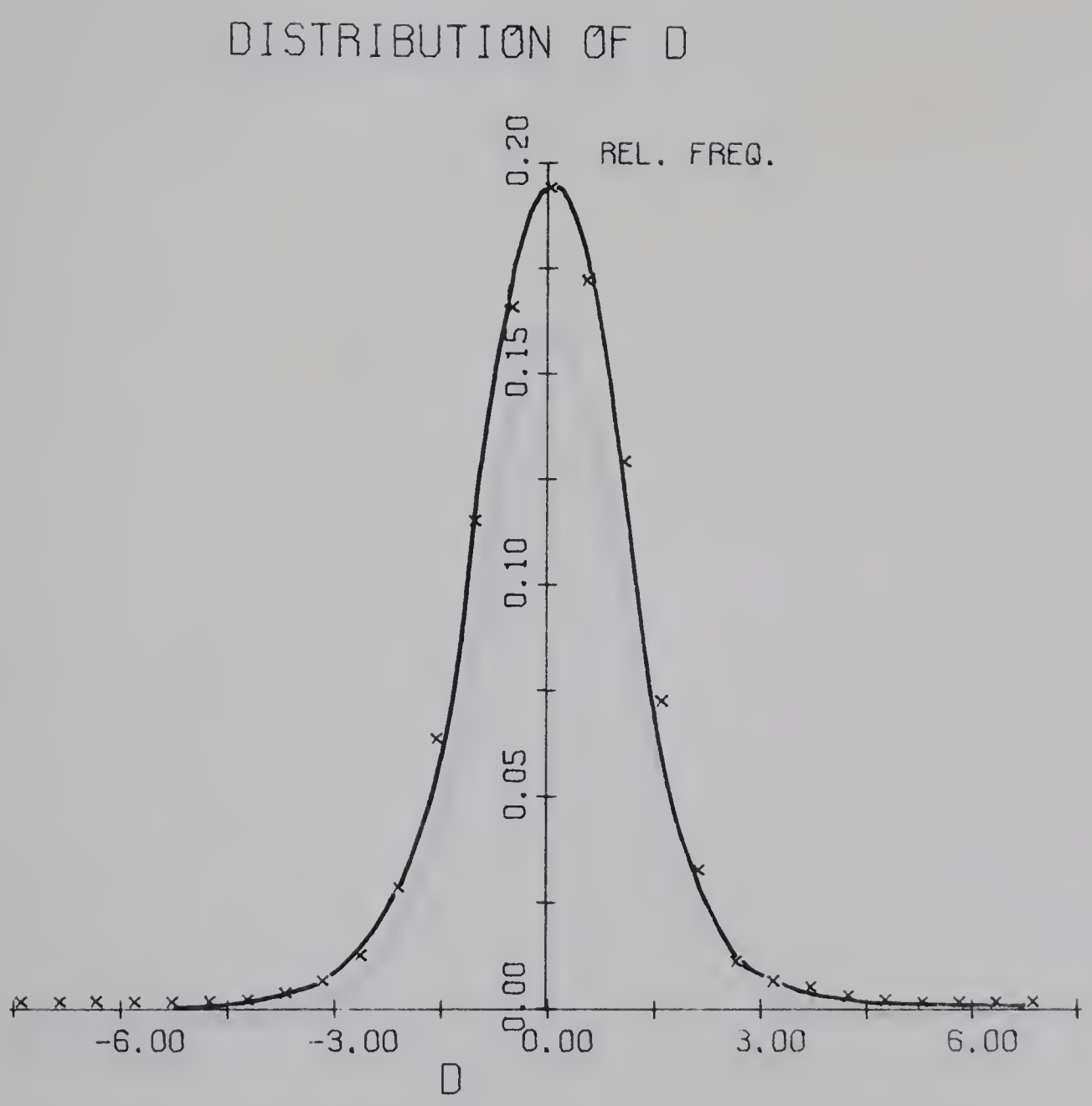


Figure 4.2. Distribution of d at $n_1 = n_2 = 8$, $\theta = 15°$.

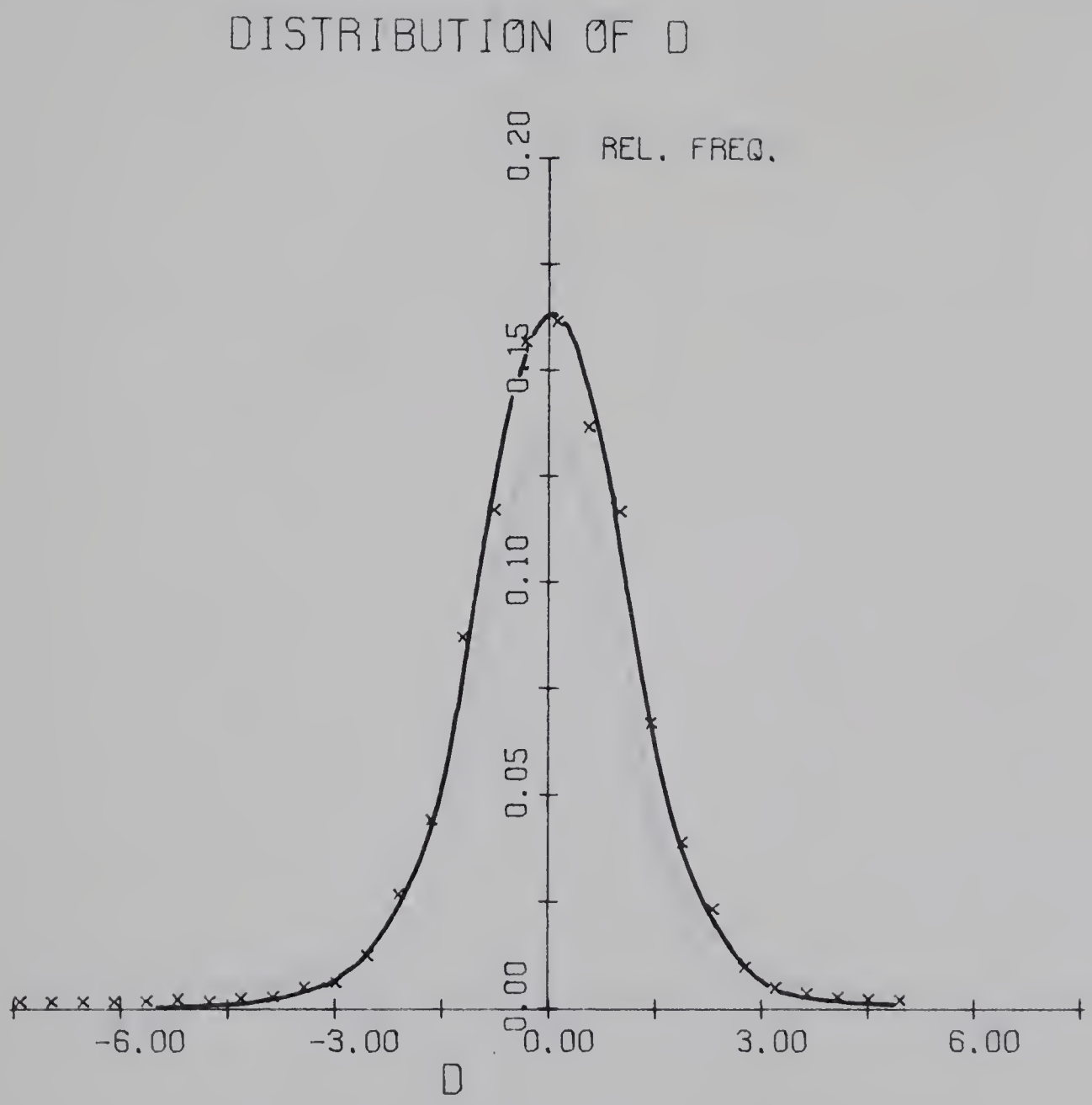
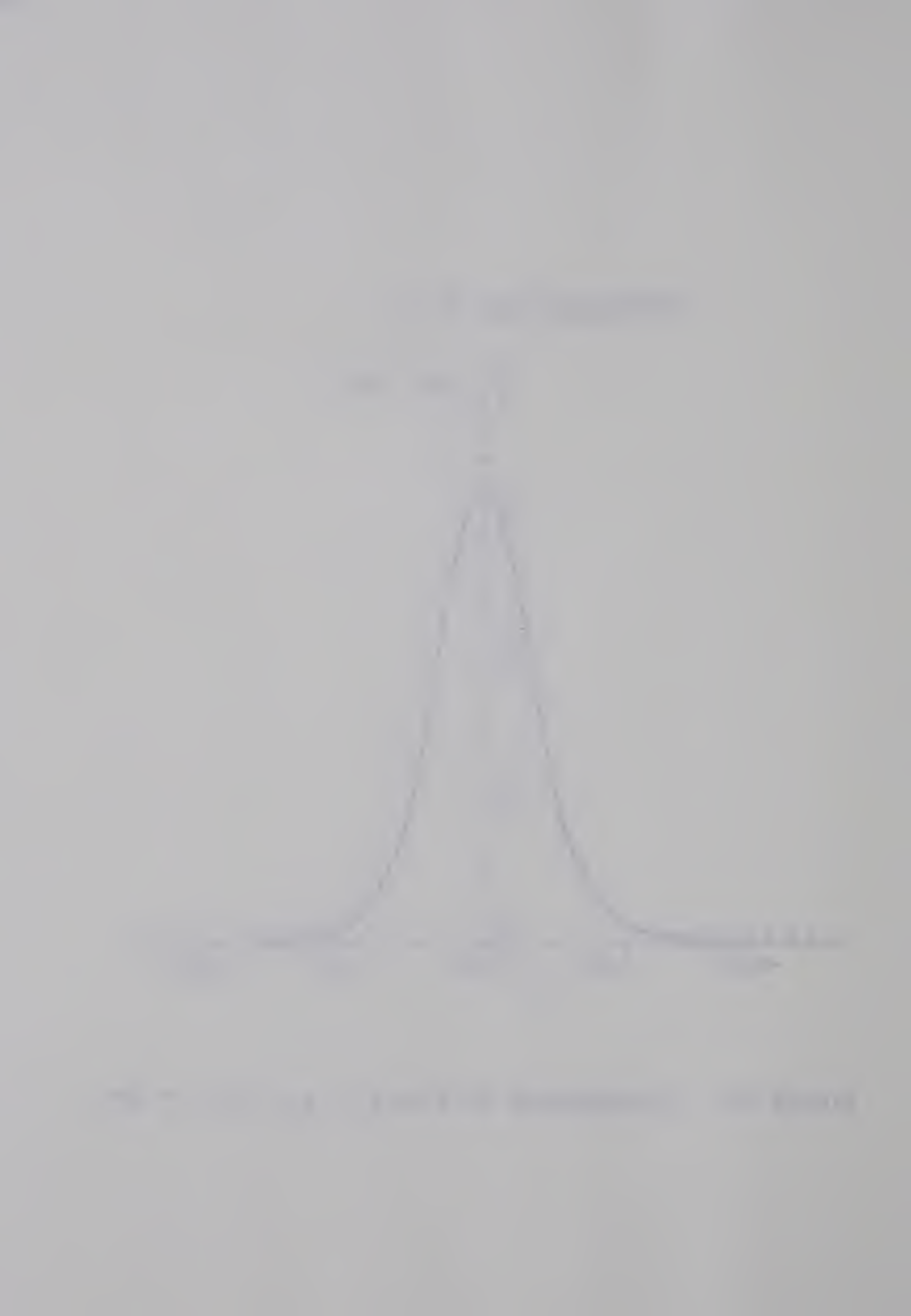


Figure 4.3. Distribution of d at $n_1 = n_2 = 8$, $\theta = 30°$.



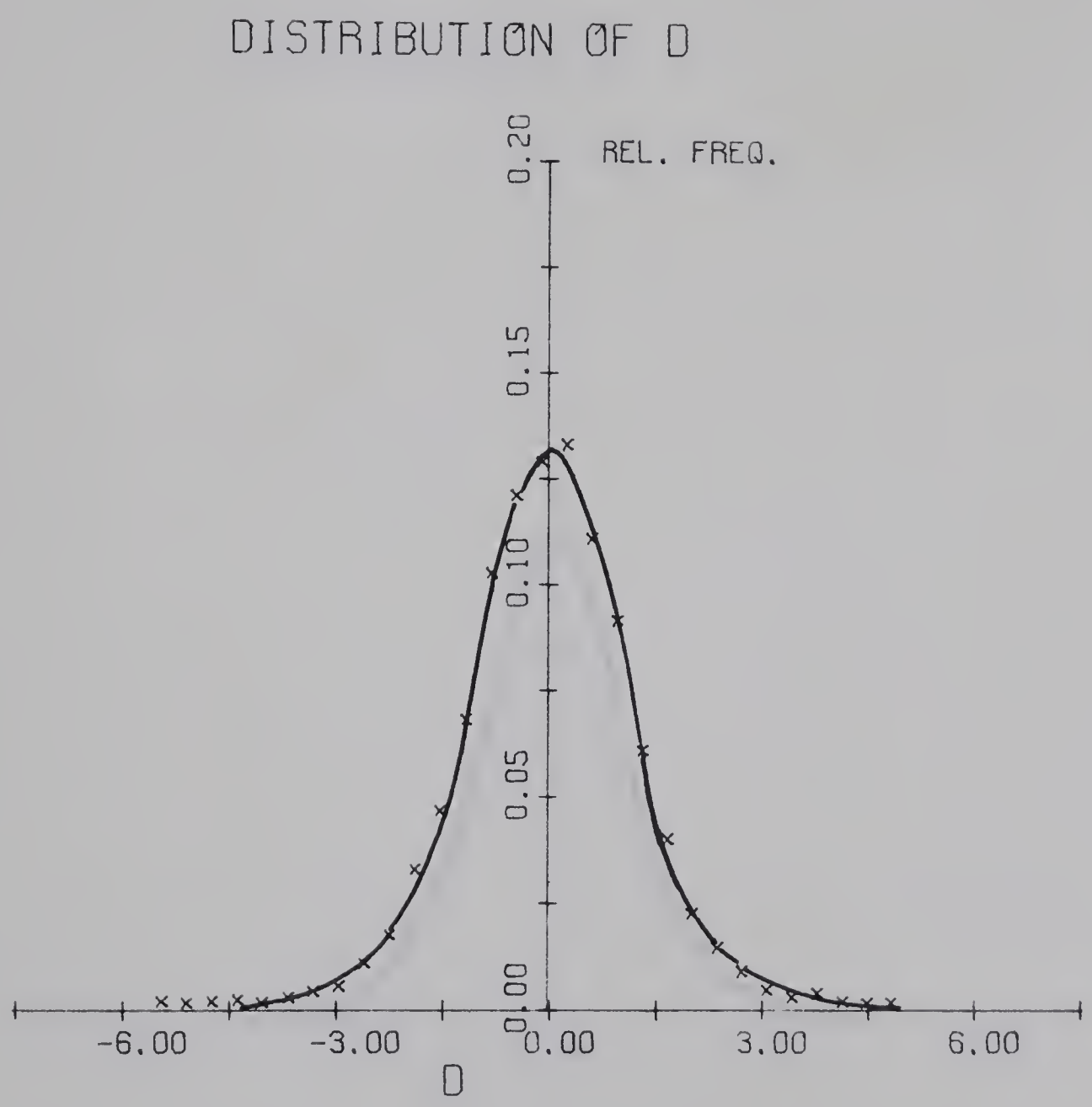
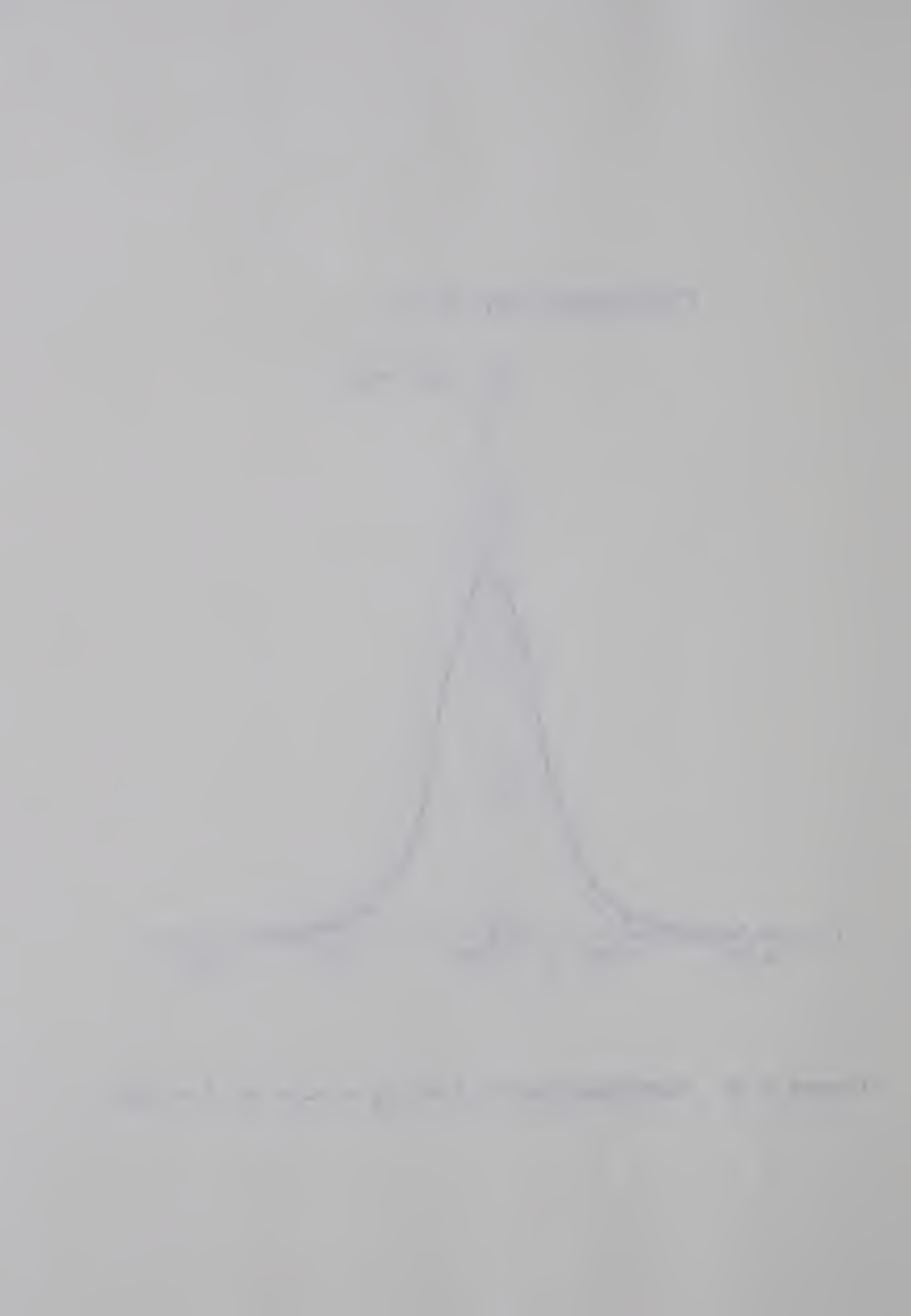


Figure 4.4. Distribution of d at $n_1 = n_2 = 8$, $\theta = 45°$.



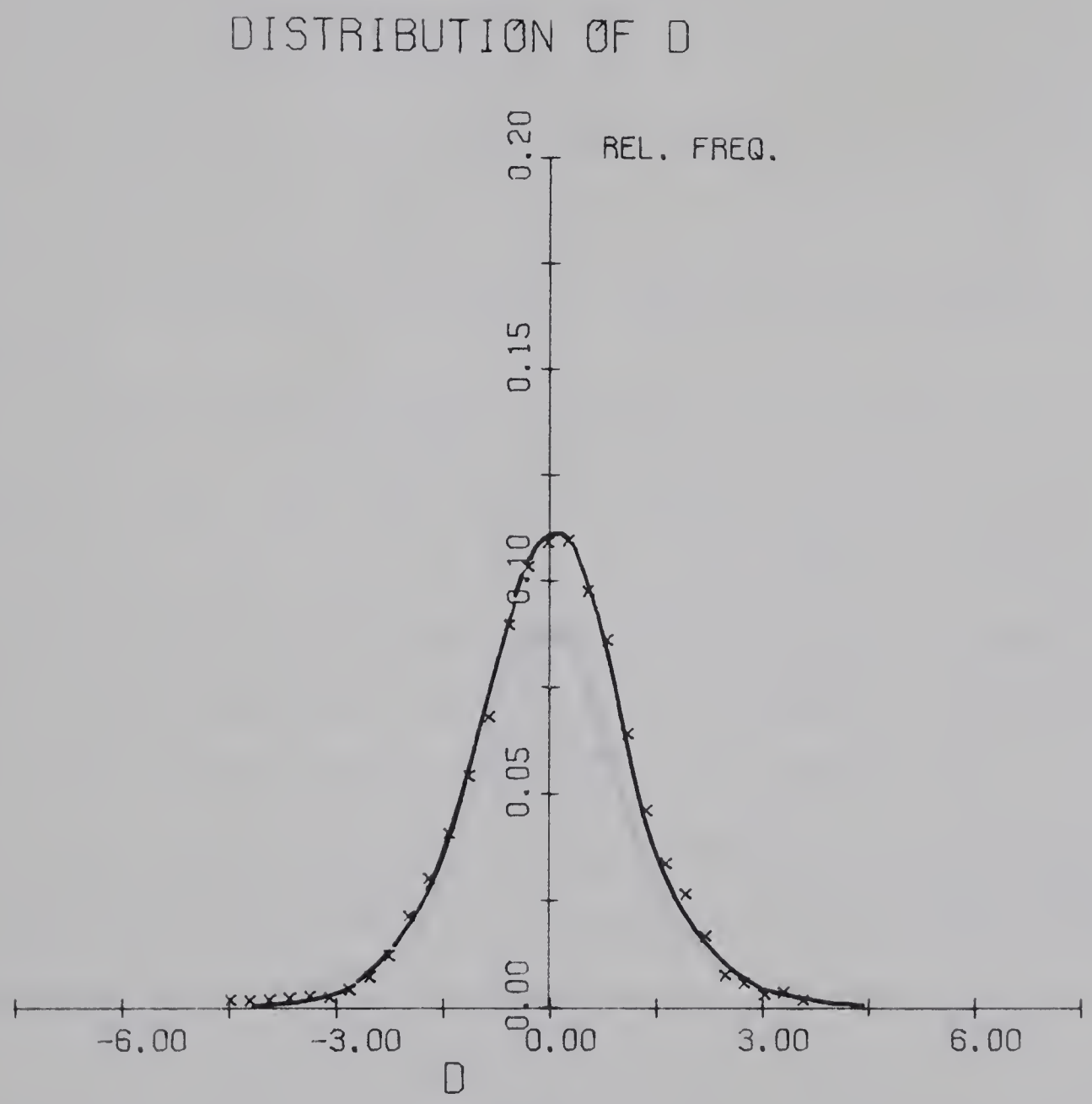
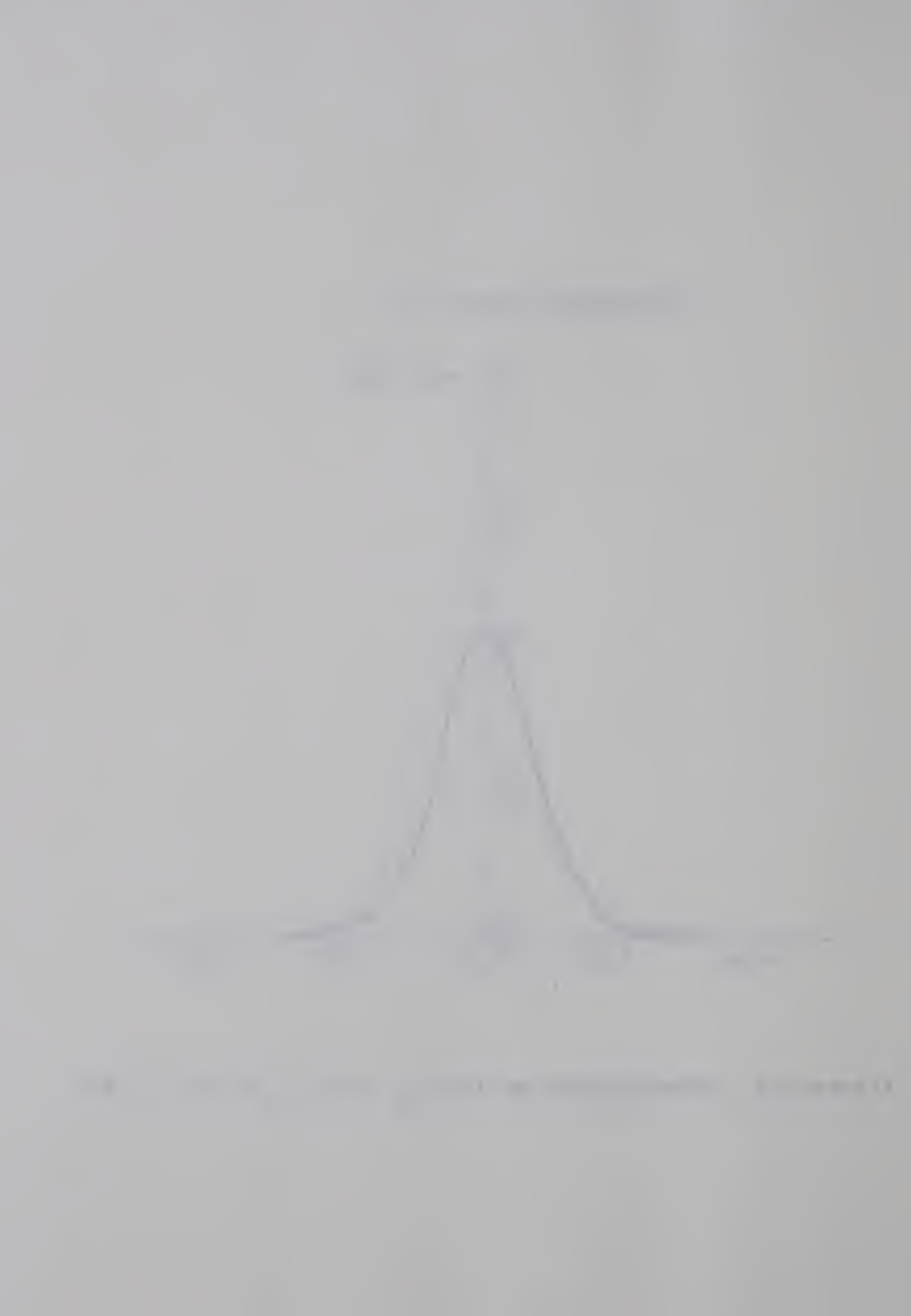


Figure 4.5. Distribution of d at $n_1 = 12$, $n_2 = 24$, $\theta = 45°$.



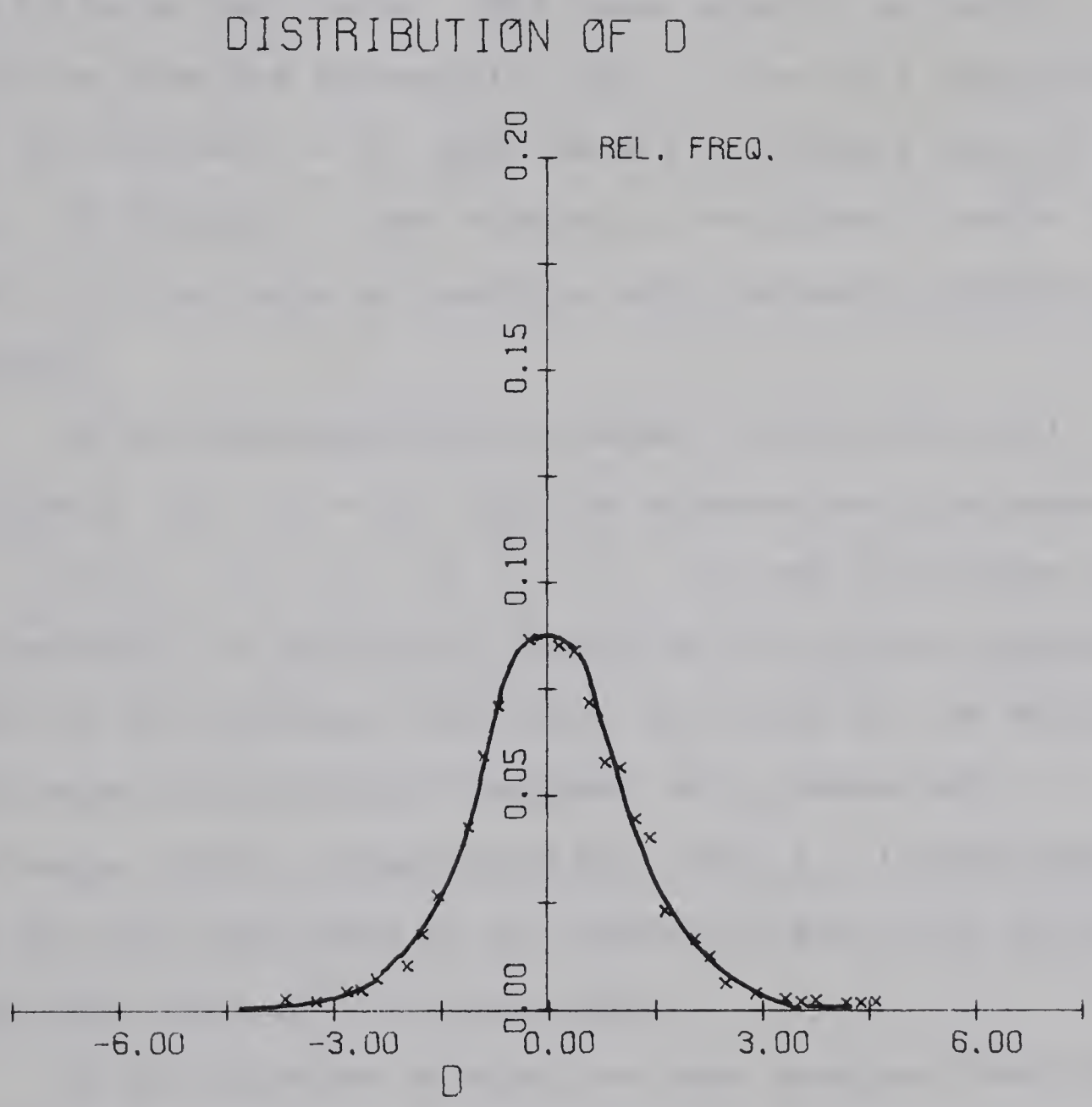
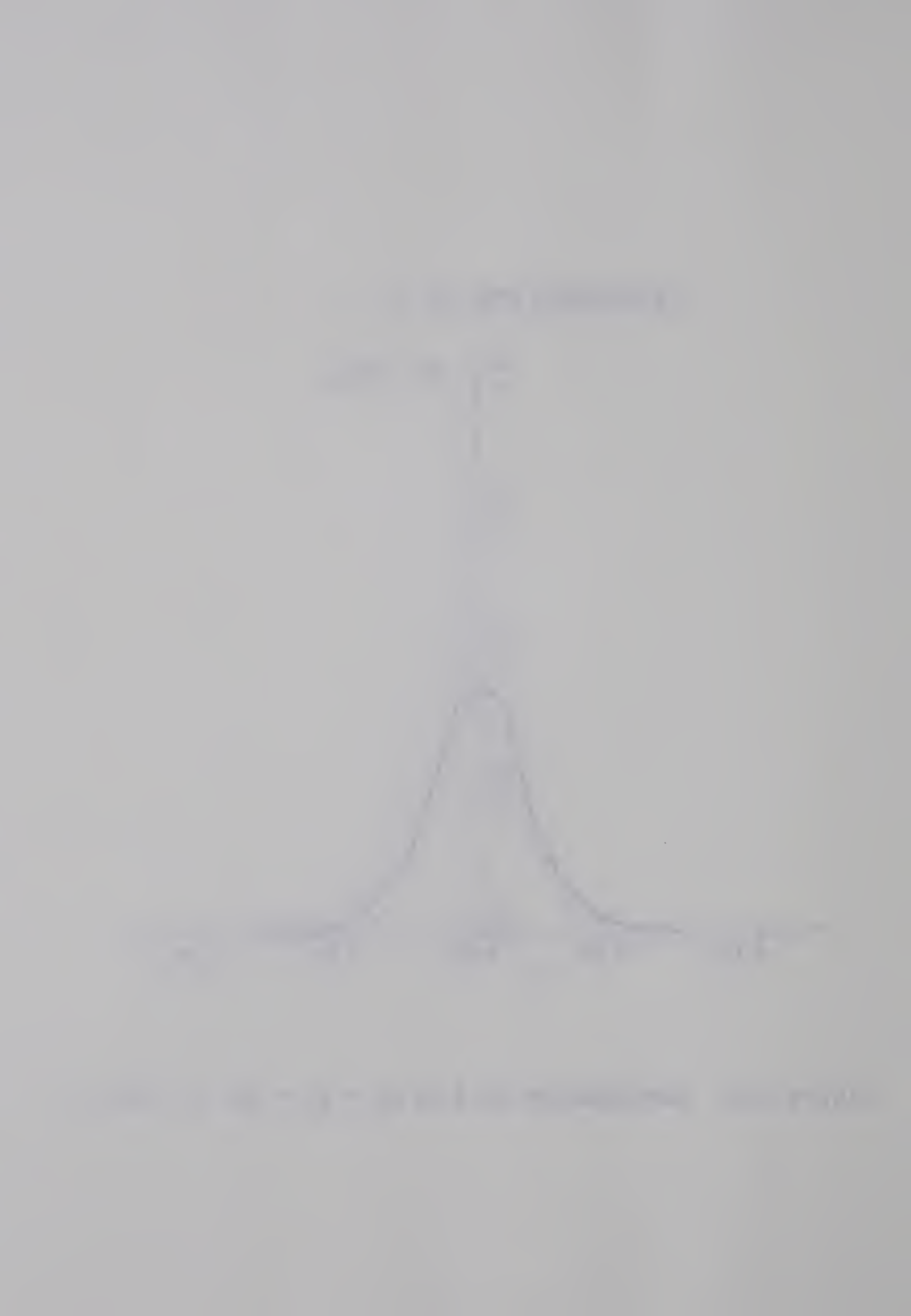


Figure 4.6. Distribution of d at $n_1 = n_2 = 24$, $\theta = 15°$.



D. Results and Discussion on the Power of the Test.

Introduction

The power of the test of hypothesis is defined as the probability that a null hypothesis will be rejected when it is in fact false. The power when no deviation exists is then the probability that a true null hypothesis will be rejected, or in other words, the actual size of the test. In choosing a test statistic, one ideally wants the power to be as large as possible when the null hypothesis is false.

In the Behrens-Fisher problem, we have the null hypothesis $H_o : \mu_1 = \mu_2$, and the alternative hypotheses $H_1 : \mu_1 \neq \mu_2$, $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$. In the literature we have covered, to decide the merits of the various proposed tests for the problem, the powers and sizes of the various tests were calculated and compared (e.g. Mehta and Srivivasan (1970), Bennett and Hsu (1961)). It was found that the size and power of the Behrens-Fisher test are much lower than those of the other tests.

In the previous section, we have examined the size of the test taking into consideration the assumption of the fiducial solution. We have found the actual size to be close to the nominal size. We expect that using the same procedure the power performance of the test would also be good. We thus use the same procedure to examine the power of the test.

Procedure

The procedure to calculate the power of the test is essentially the same as that for calculating the actual size except that in the former the true difference in the population means is nonzero. The same samples generated to calculate the actual size are thus used to calculate the power of the test. The only difference is in the numerator of the d statistic which is now normally distributed with a nonzero mean equal to $-\mu_2$.

The alternative hypothesis used in our calculations is $H_1 : \mu_1 < \mu_2$. For each set of parameters (n_1, n_2, θ) and $\mu_1 = 0$, μ_2 is allowed to vary from 1 to 6. The power for each of the size specifications was calculated from the same sample as the relative frequency that the observed d values exceed the respective critical values. That is,

power of the test = rel. frequency $\left\{|d_{obs}| \geq d_\alpha | \mu_1 < \mu_2\right\}$

The empirical distribution of d under the alternative hypothesis $\mu_2 - \mu_1 = 1$, $\mu_2 - \mu_1 = 5$ was also obtained and their plots were drawn for several cases.

Results

Some typical results on the power of the test for some sets of parameters (n_1, n_2, θ) are tabulated in Table 4.11. Power curves for some results are drawn and presented in Figures 4.7 - 4.10. Curves for the empirical distributions of d are presented in Figures 4.11 - 4.14.

Table 4.11. Power of the Behrens-Fisher Test for Various δ, where $\delta = \mu_2 - \mu_1$ and the Nominal Size = α.

(a) $n_1 = n_2 = 3$

	α \ δ	0.0	1.0	2.0	3.0	4.0	5.0	6.0
$\theta = 15°$	.10	.101	.121	.241	.385	.520	.653	.757
	.05	.046	.057	.128	.234	.361	.476	.595
	.02	.017	.018	.048	.100	.174	.265	.366
	.01	.009	.010	.022	.047	.088	.144	.217

	α \ δ	0.0	1.0	2.0	3.0	4.0	5.0	6.0
$\theta = 30°$	.10	.101	.219	.484	.720	.864	.934	.970
	.05	.048	.119	.330	.558	.745	.868	.934
	.02	.019	.048	.048	.164	.345	.529	.700
	.01	.008	.024	.088	.207	.361	.519	.664

	α \ δ	0.0	1.0	2.0	3.0	4.0	5.0	6.0
$\theta = 45°$	.10	.102	.310	.686	.898	.971	.991	.997
	.05	.047	.188	.514	.783	.925	.976	.992
	.02	.018	.086	.300	.570	.781	.915	.966
	.01	.008	.041	.185	.404	.617	.790	.906

(cont'd.)

Table 4.11 (cont'd.)

(b) $n_1 = 6$, $n_2 = 12$

	α \ δ	0.0	1.0	2.0	3.0	4.0	5.0	6.0
θ = 15°	.05	.050	.111	.292	.529	.730	.871	.942
	.01	.010	.032	.116	.285	.487	.677	.821

	α \ δ	0.0	1.0	2.0	3.0	4.0	5.0	6.0
θ = 30°	.05	.052	.259	.697	.932	.987	.999	1.0
	.01	.009	.100	.434	.794	.949	.989	.999

	α \ δ	0.0	1.0	2.0	3.0	4.0	5.0	6.0
θ = 45°	.05	.051	.408	.885	.992	1.0	1.0	1.0
	.01	.010	.181	.697	.952	.996	1.0	1.0

	α \ δ	0.0	1.0	2.0	3.0	4.0	5.0	6.0
θ = 60°	.05	.052	.521	.970	.999	1.0	1.0	1.0
	.01	.010	.242	.826	.990	1.0	1.0	1.0

	α \ δ	0.0	1.0	2.0	3.0	4.0	5.0	6.0
θ = 75°	.05	.050	.583	.989	1.0	1.0	1.0	1.0
	.01	.010	.261	.881	.998	1.0	1.0	1.0

Discussion

Table 4.11 and Figures 4.7 - 4.10 may be summarized as follows:

(i) For a given set (n_1, n_2, θ) the power increases for increasing test size.

(ii) For fixed θ and size, the power increases with increases in n_1 or n_2.

(iii) For fixed n_1, n_2 and the size of the test, the power increases with increasing sample variance ratio.

These results are calculated with the assumptions of the fiducial solution. A direct comparison with the results in Mehta and Srinivasan (1970) and Bennett and Hsu (1961) is not possible since they used the confidence interval approach and calculated the power with respect to n_1, n_2 and true variance ratio. However, since the actual size of the test we calculated is close to the nominal size, the power is also raised correspondingly. We thus say that the power performance of the test is acceptable for the practical user.

Figures 4.11 - 4.14 show that when the null hypothesis is not true, i.e. $\mu_2 - \mu_1 > 0$, the distribution of d becomes more and more shifted to the left as the difference between the true means increases and also as θ increases.

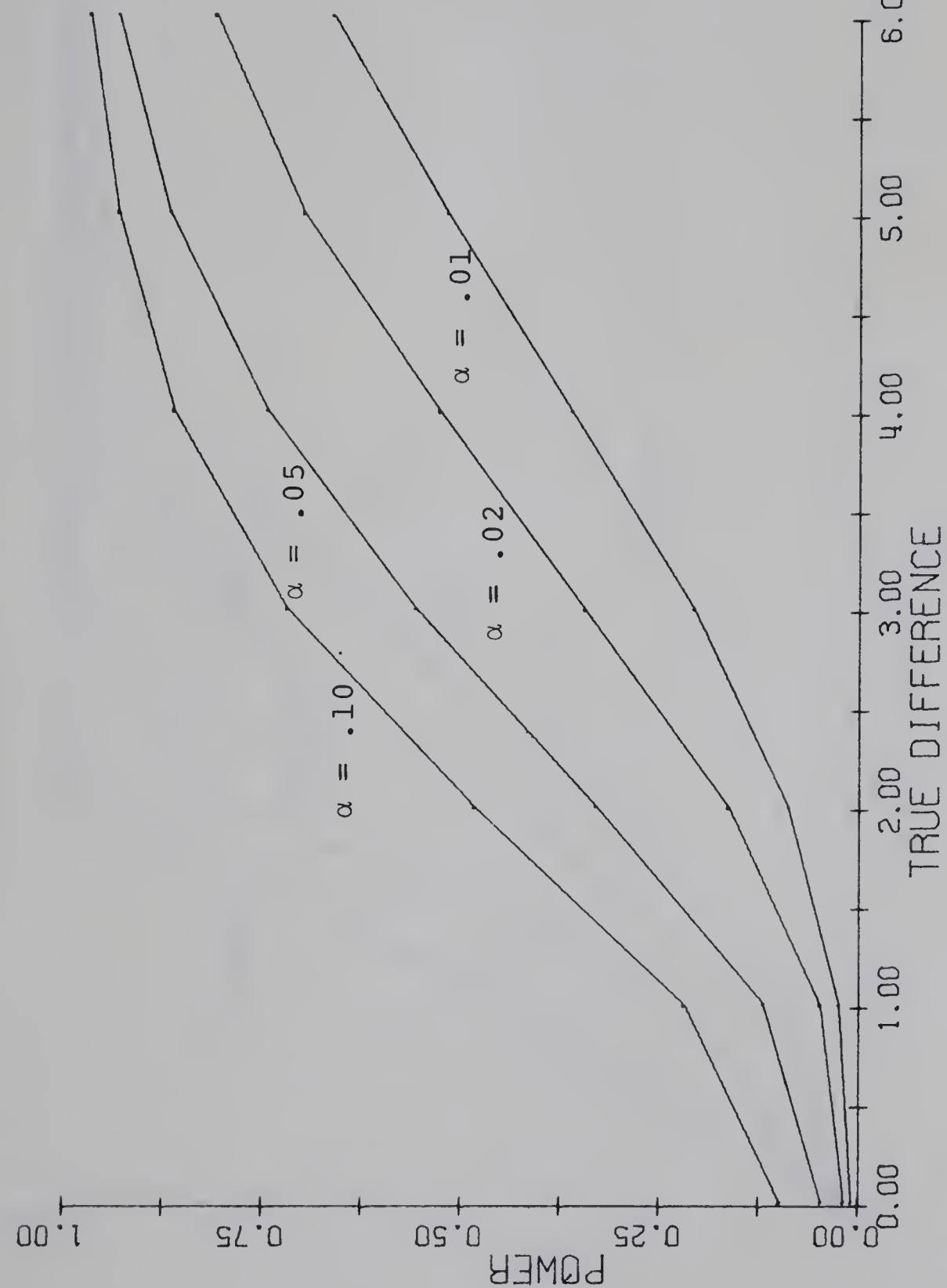


Figure 4.7. Typical Empirical Power Curves of The Behrens-Fisher Test for Fixed Sample Sizes, θ and Various Nominal Test Sizes.

$n_1 = n_2 = 3$, $\theta = 30°$.

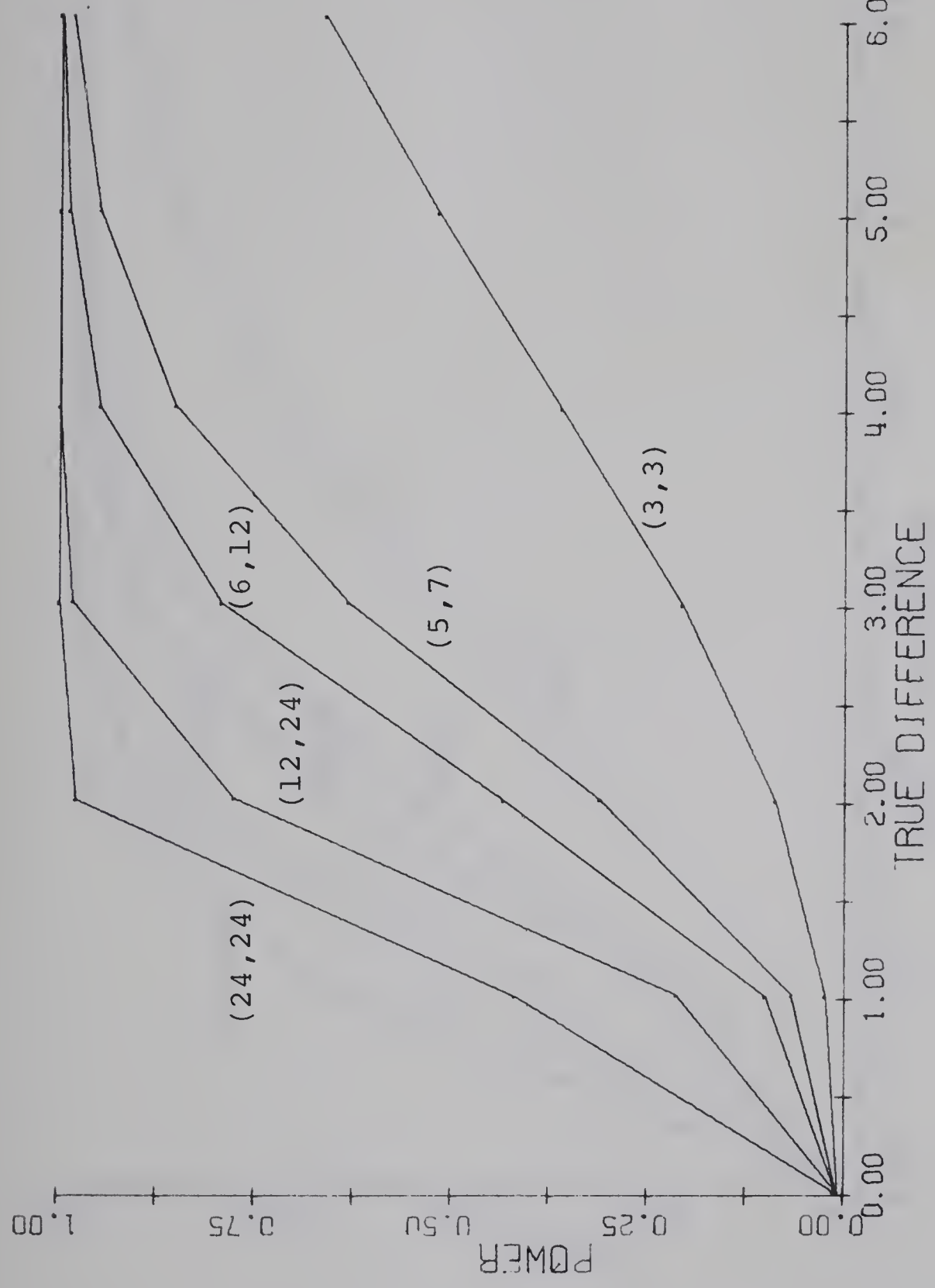


Figure 4.8. Typical Empirical Power Curves of the Behrens-Fisher Test for Fixed Nominal Size and θ and Various (n_1, n_2). Nominal Size = .01, $\theta = 30°$.

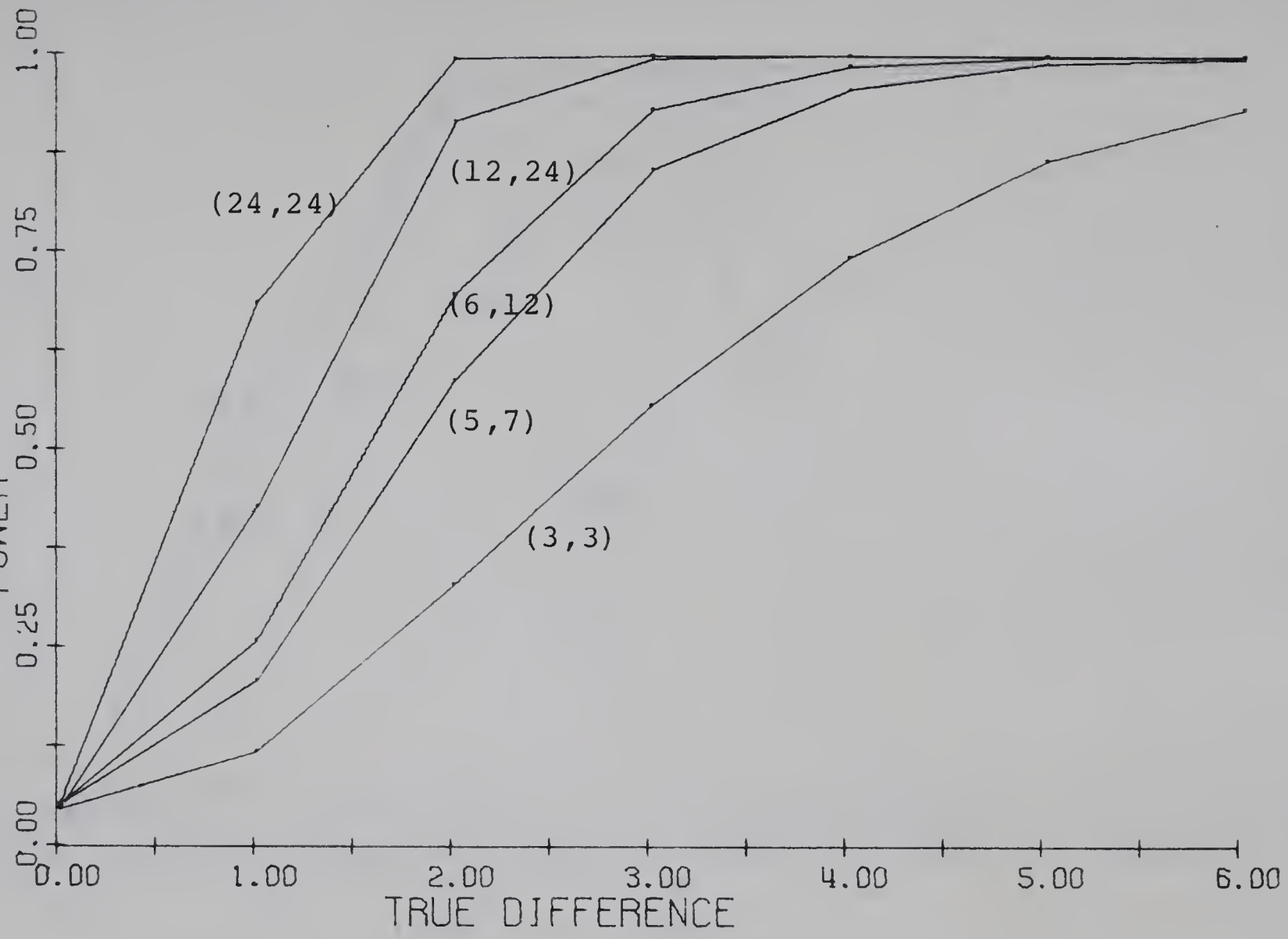


Figure 4.9. Typical Empirical Power Curves of the Behrens-Fisher Test for Fixed Nominal Size and θ and Various (n_1, n_2). Nominal Size = .05, $\theta = 30°$.

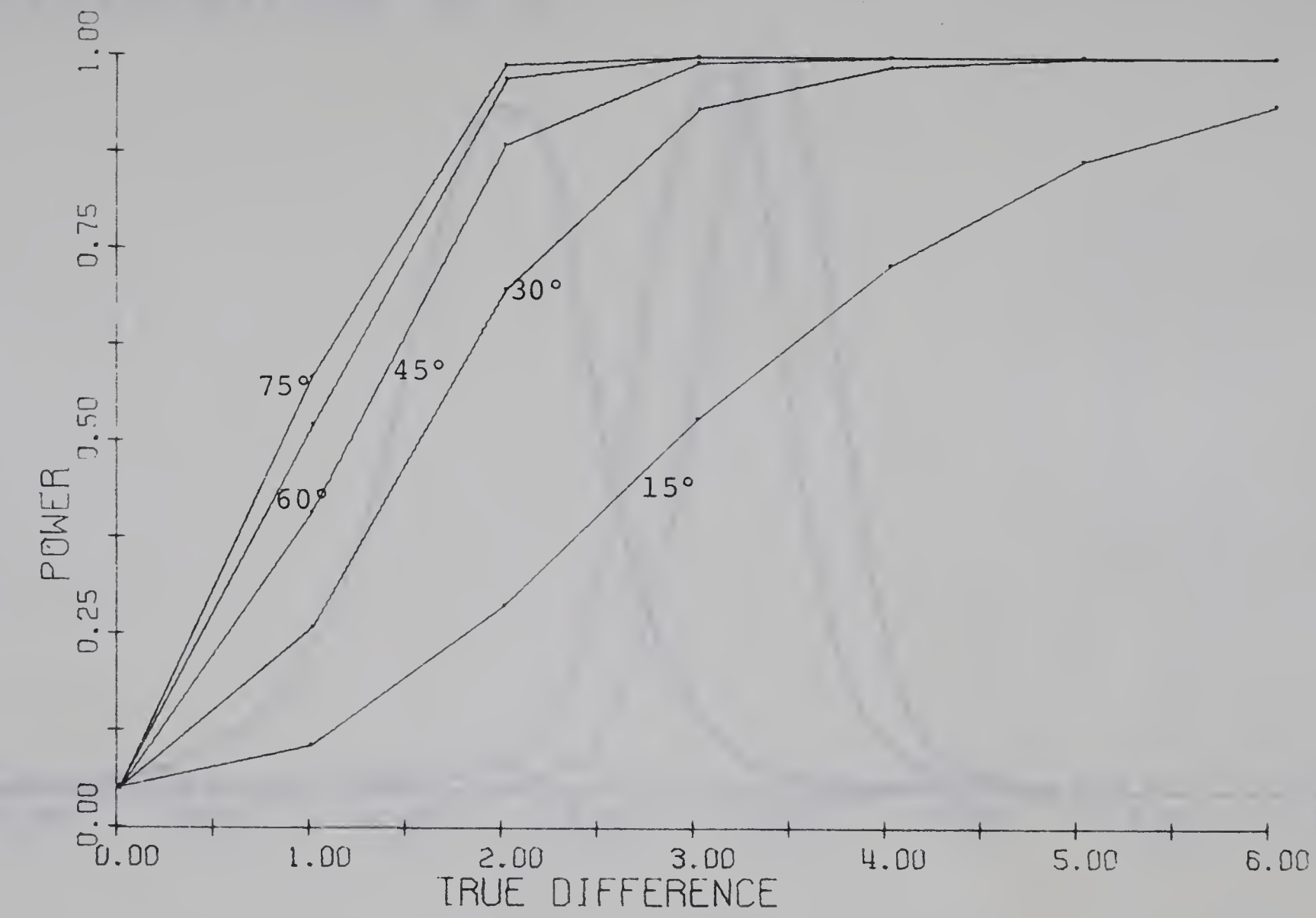


Figure 4.10. Typical Empirical Power Curves of the Behrens-Fisher Test for Fixed Sample Size and Nominal Test Size and Various θ.
$n_1 = 6$ $n_2 = 12$, Nominal Size = .05.

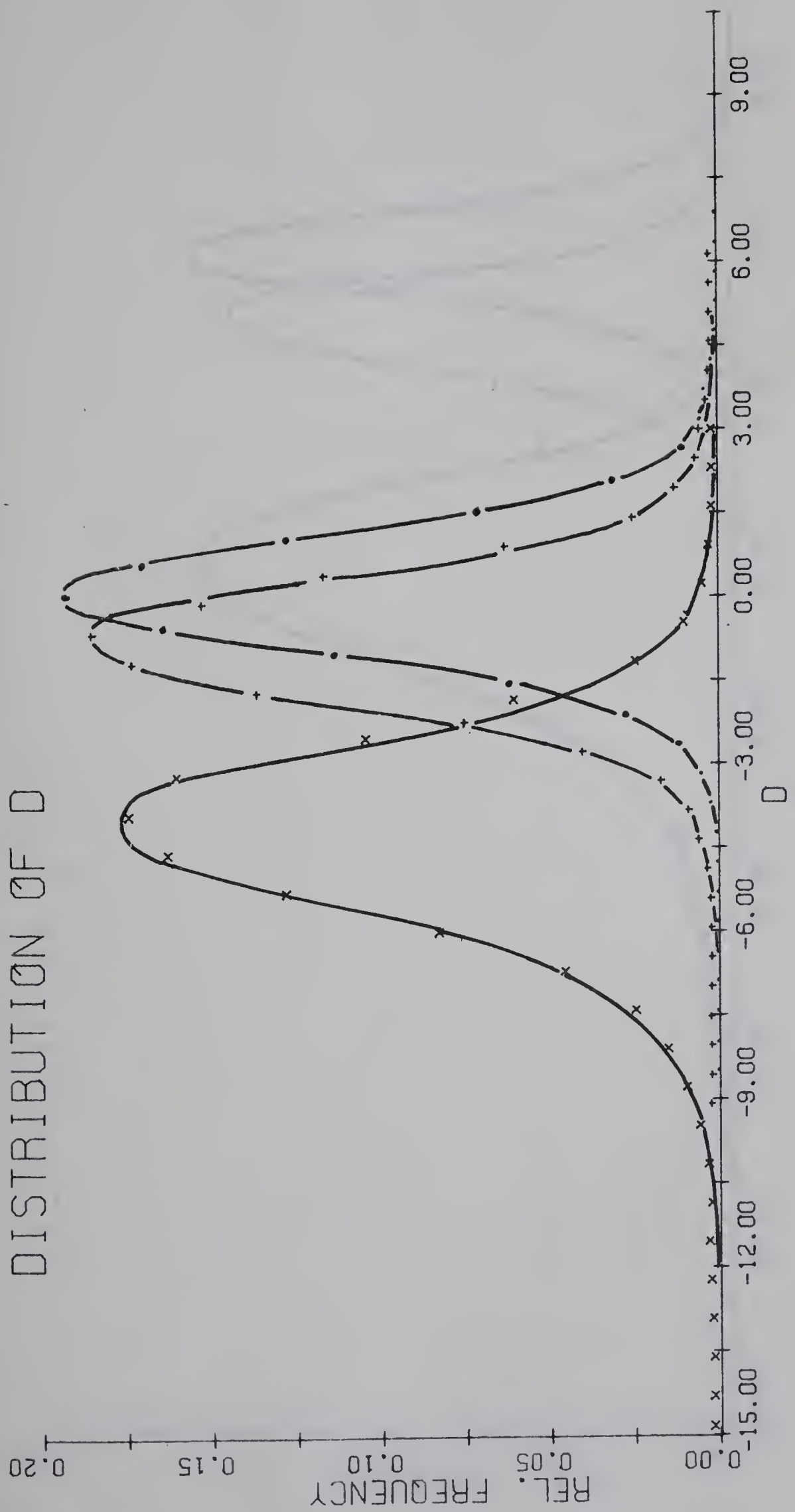
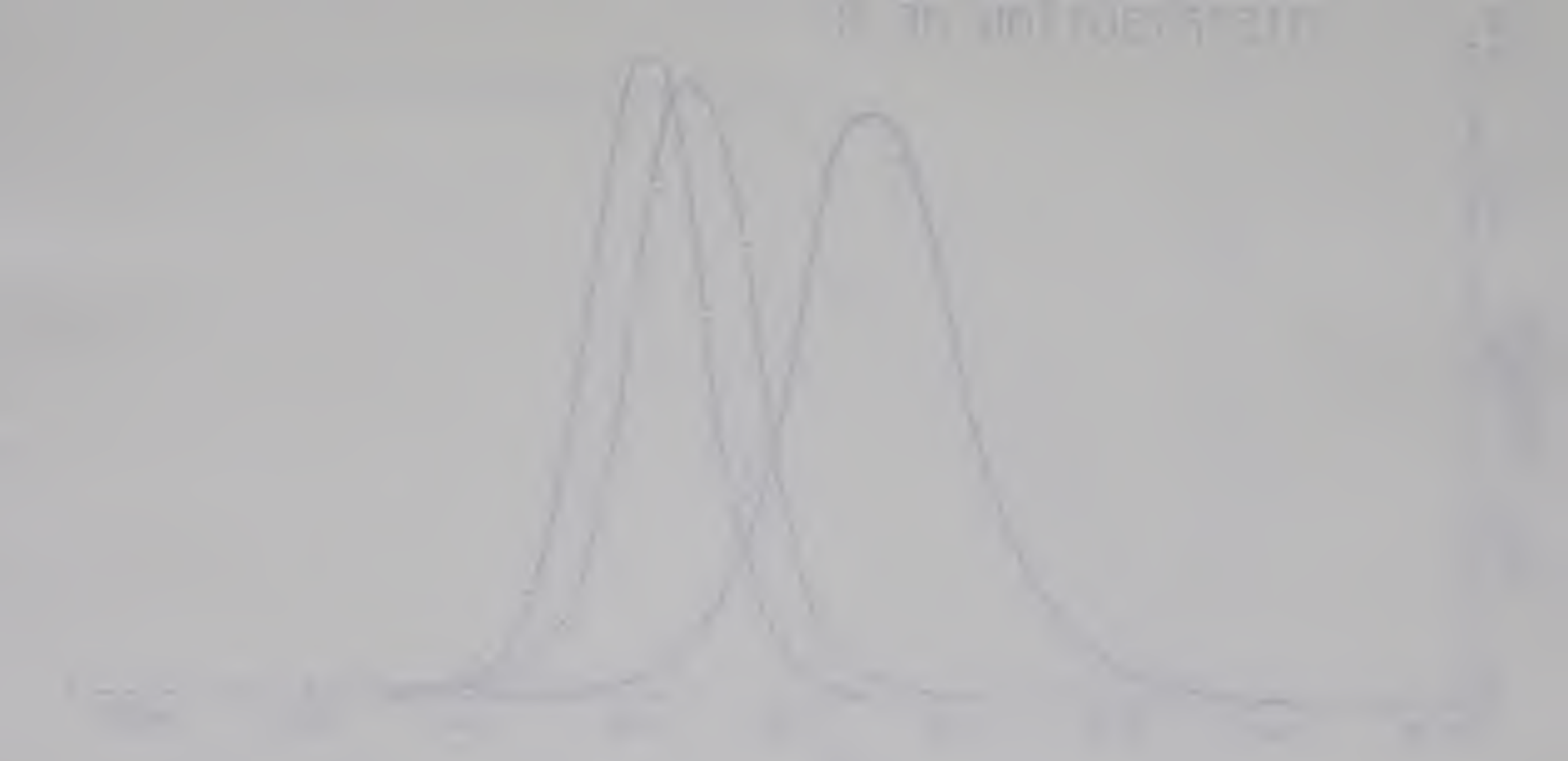


Figure 4.11. The Empirical Distribution of the d Statistic Under the Null Hypothesis $\mu_2 - \mu_1 = 0$, and the Alternative Hypothesis $\mu_2 - \mu_1 = 1$, $\mu_2 - \mu_1 = 5$ ($n_1 = 8$, $n_2 = 8$, $\theta = 15°$)



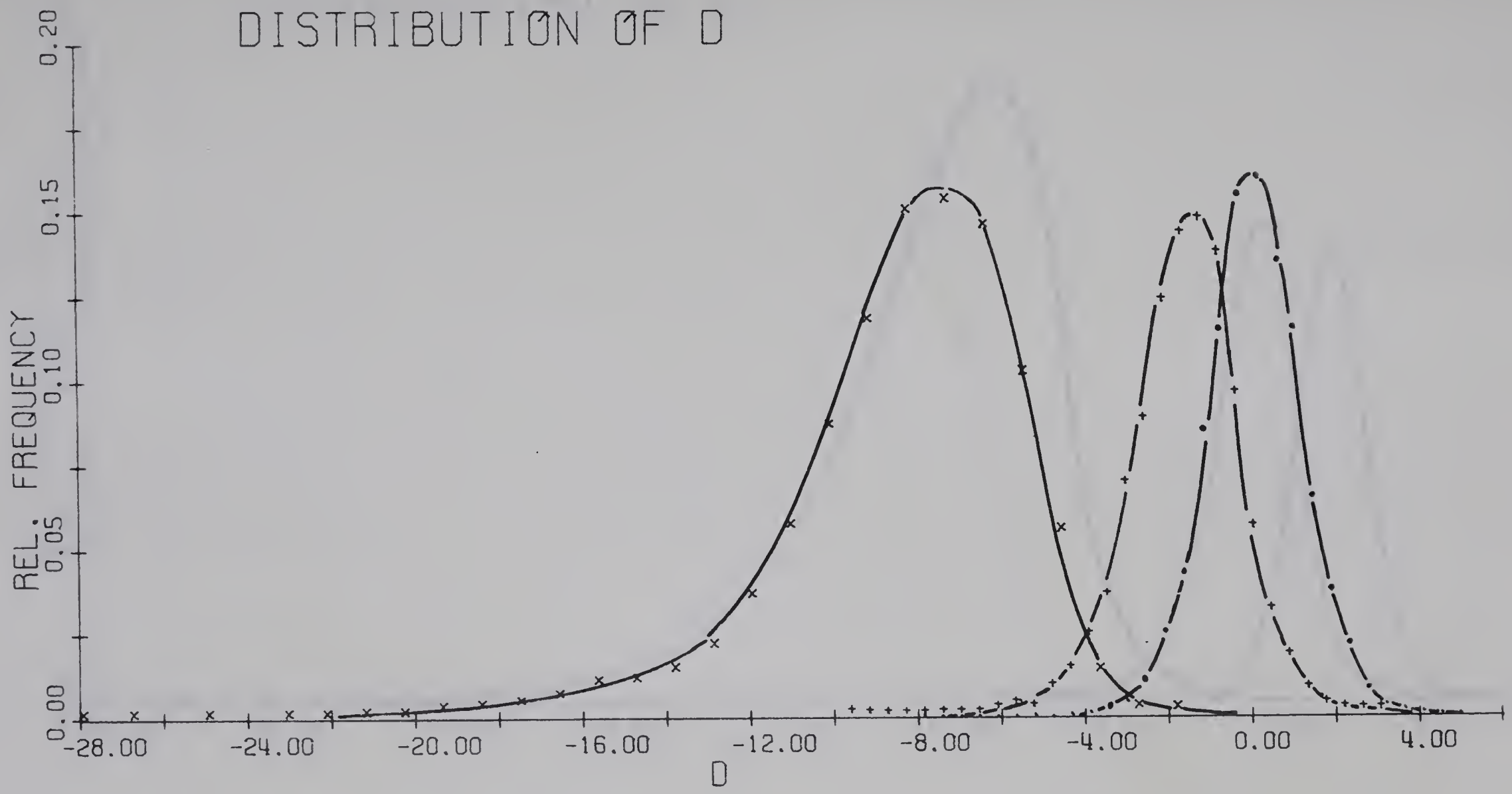


Figure 4.12. The Empirical Distribution of the d Statistic Under the Null Hypothesis $\mu_2 - \mu_1 = 0$, and the Alternative Hypothesis $\mu_2 - \mu_1 = 1$, $\mu_2 - \mu_1 = 5$ ($n_1 = 8$, $n_2 = 8$, $\theta = 30°$)

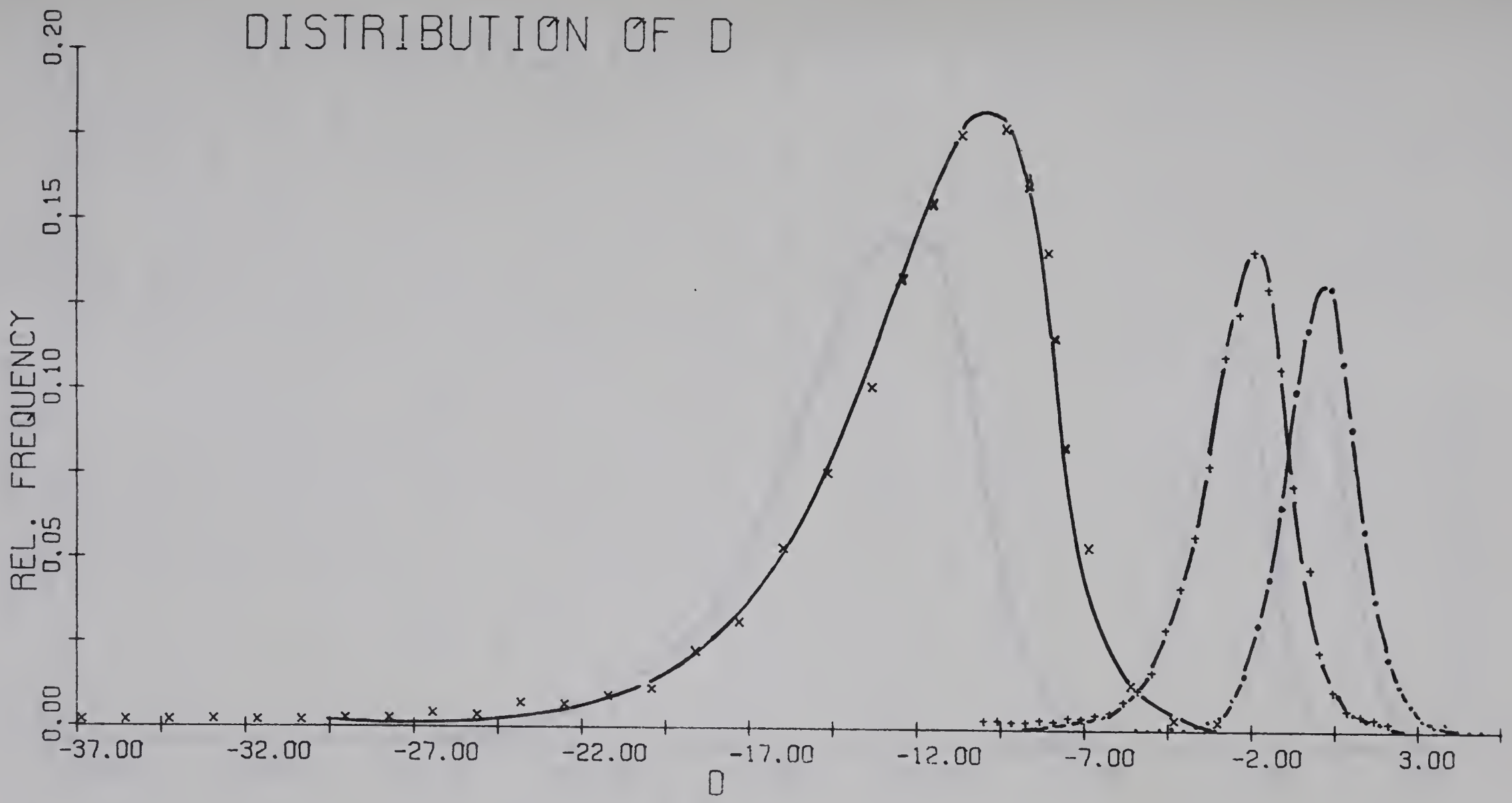


Figure 4.13. The Empirical Distribution of the d Statistic Under the Null Hypothesis $\mu_2 - \mu_1 = 0$, and the Alternative Hypothesis $\mu_2 - \mu_1 = 1$, $\mu_2 - \mu_1 = 5$ ($n_1 = 8$, $n_2 = 8$, $\theta = 45°$)

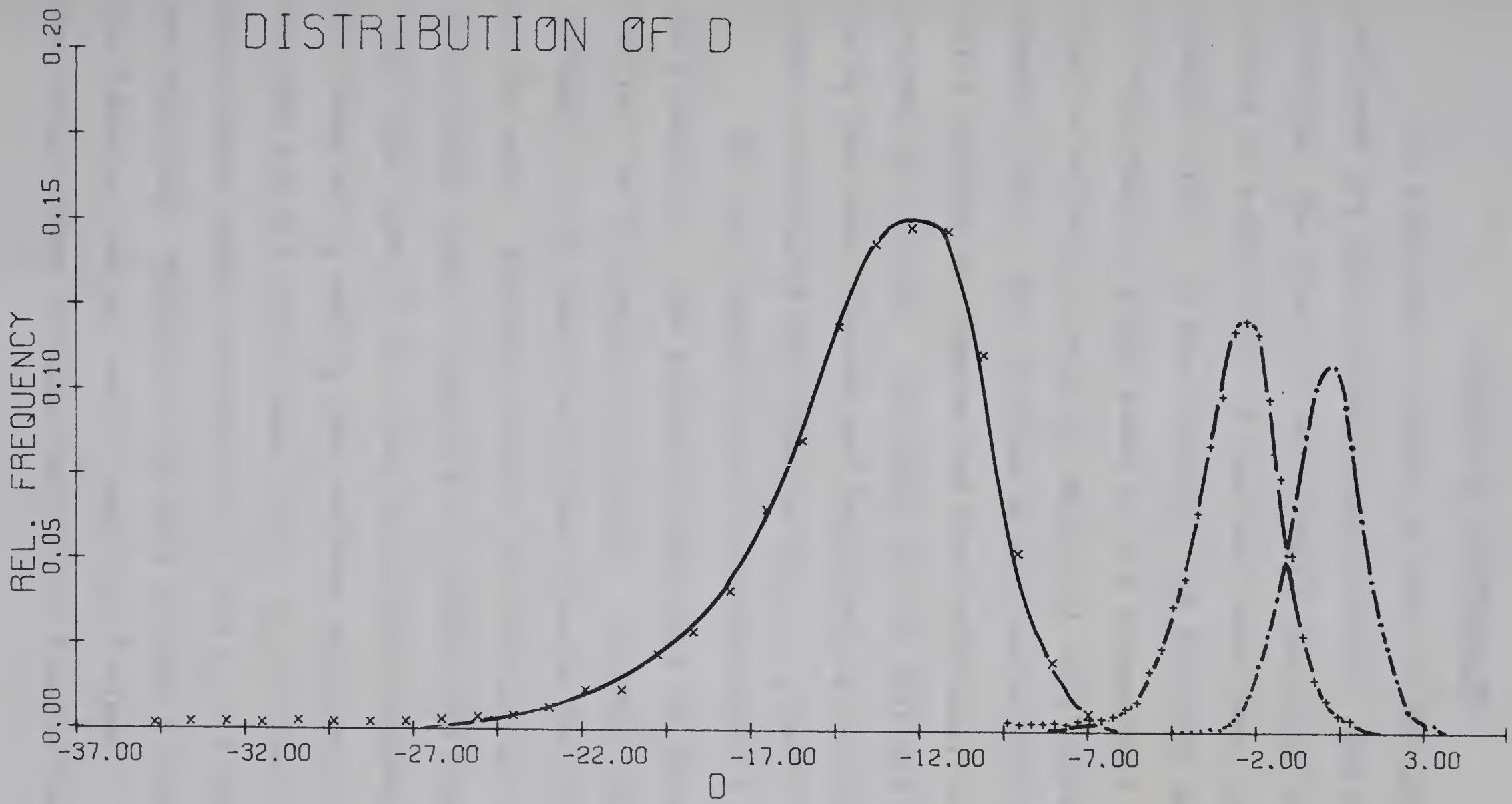


Figure 4.14. The Empirical Distribution of the d Statistic Under the Null Hypothesis $\mu_2 - \mu_1 = 0$, and the Alternative Hypothesis $\mu_2 - \mu_1 = 1$, $\mu_2 - \mu_1 = 5$ ($n_1 = 24$ $n_2 = 24$, $\theta = 45°$)

CHAPTER V. CONCLUSION

We have made a survey of some of the major solutions proposed for the Behrens-Fisher problem. The Behrens-Fisher solution, the first to be proposed, has been criticized for giving an actual size of the test much lower than the nominal size. We have examined the solution and we showed in our sampling study based on the assumptions of the test that it actually yields an empirical size close to the nominal size. This conclusion applies to both small and large degrees of freedom and more especially to larger degrees of freedom. The power of the test was studied using the same approach and observations were made on the power performance with changes in n_1, n_2 and θ.

We thus consider that the objectives of this study have been met. The Behrens-Fisher test is recommended for use for the following two reasons. Firstly, the test actually yields empirical sizes close to the size specified by the user. Secondly, tables for its use are available for a wider range of degrees of freedom and sample variance ratio than some of the other solutions proposed. Based on the range of n_1 and n_2 that we have verified, the test can be used for the whole range of n_1, n_2 covered in the available tables except perhaps for n_1 and $n_2 < 3$ which we have not verified. Moreover, the test is most suitable for use for large n_1 and n_2, say n_1 and $n_2 \geq 6$ since on the average the actual sizes at these degrees of freedom have a smaller

deviation from the nominal size.

On the method of calculating the actual size, our procedure could certainly be improved by say narrowing the tolerance limit ε in obtaining samples with the correct S_1/S_2 ratio and increasing further the number of pairs of samples taken. This would be expensive to do but would increase the accuracy of the results.

Calculation and tabulation of critical values of d for more degrees of freedom would be helpful for the user of the test.

On the other hand, the general appearance of the empirical d distribution in Figures 4.1 - 4.6 in Chapter IV, and the approximation of Patil (1964) suggests the possibility of approximating the d distribution by the t distribution. Hence another aspect of the test that is worth looking into is the possibility of using the t tables which are more readily available for the critical values of d. One could investigate the approximation at various degrees of freedom and sample variance ratios. If the approximation is good, then the Behrens-Fisher test will be even more available for practical use.

VI. BIBLIOGRAPHY

Abramowitz, M. and Stegun, I.A. (1964). Handbook of Mathematical Functions, NBS Applied Mathematical Series, Vol. 55, Washington, D.C.

Anderson, R.L. (1942). "Distribution of the Serial Correlation Coefficient," Annals of Mathematical Statistics, Vol. 13, p.1.

Aspin, A.A. (1948). "An Examination and Further Development of a Formula Arising in the Problem of Comparing Two Mean Values," Annals of Mathematical Statistics, Vol. 35, pp. 88-96.

Aspin, A.A. (1949). "Tables for Use in Comparison Whose Accuracy Involves Two Variances, Separately Estimated, with Appendix by B.L. Welch," Biometrika, Vol. 36, pp. 290-296.

Bartlett, M.S. (1936). "The Information Available in Small Samples," Cambridge Philosophical Society Proceedings, Vol. 32, pp. 560-566.

Behrens, W.U. (1929). "Ein Beitrag Zur Fehlerberechung Bei Wenigen Beobachtungen," Landwirtschaftliche Jahrbucher, Vol. 68, pp. 807-837.

Behrens, W.U. (1964). "The Comparison of Means of Independent Normal Distributions with Different Variances," Biometrics, Vol. 20, pp. 17-27.

Bennett, B.M. and Hsu, P. (1961). "Sampling Studies on the Behrens-Fisher Problem," Metrika, Vol. 4, pp. 89-104.

Box, G.E.P. and Muller, M.E. (1958). "A Note on the Generation of Random Normal Deviates," Annals of Mathematical Statistics, Vol. 29, pp. 610-611.

Box, G.E.P. and Tiao, G.C. (1973). Bayesian Inference in Statistical Analysis, Addison-Wesley Publishing Company, Massachusetts.

Chen, E.H. (1971). "A Random Normal Number Generator for 32-Bit Word Computers," Journal of the American Statistical Association, Vol. 66, pp. 400-403.

Cornish, E.A. and Fisher, R.A. (1937). "Moments and Cumulants in the Specifications of Distributions," Extrait de la Revue de l'Institut International de Statistique, Vol. 4, pp. 1-14.

Downham, D.Y. and Roberts, F.D.K. (1967). "Multiplicative Congruential Pseudo-Random Number Generators," Computer Journal, Vol. 10, pp. 74-77.

Fisher, R.A. (1935). "The Fiducial Argument in Statistical Inference," Annals of Eugenics, Vol. 6, pp. 391-398.

Fisher, R.A. (1937). "On a Point Raised by M.S. Bartlett on Fiducial Probability," Annals of Eugenics, Vol. 7, pp. 370-375.

Fisher, R.A. (1939). "Comparison of Samples with Possibly Unequal Variances," Annals of Eugenics, Vol. 9, pp. 174-180.

Fisher, R.A. (1941). "The Asymptotic Approach to Behrens' Integral with Further Tables for the d Test of Significance," Annals of Eugenics, Vol. 11, pp. 141-172.

Fisher, R.A. (1956). "On a Test of Significance in Pearson's Biometrika Tables," Journal of the Royal Statistical Society B, Vol. 18, pp. 36-40.

Fisher, R.A. and Healy, M.J.R. (1956). "New Tables of Behrens' Test of Significance," Journal of the Royal Statistical Society, Vol. 18, pp. 212-216.

Fisher, R.A. (1959). Statistical Methods and Scientific Inference, Oliver and Boyd, Edinburgh.

Fisher, R.A. (1961). "Sampling the Reference Set," Sankhya, Vol. 23, pp. 3-8.

Fisher, R.A. and Yates, F. (1963). Statistical Tables for Biological, Agricultural and Medical Research, Oliver and Boyd, Edinburgh.

Hastings, C. (1955). Approximations for Digital Computers, Princeton University Press, Princeton.

Hill, I.D. and Pike, M.C. (1967). "Algorithm 299 Chi Squared Integral [515]," Communications of the ACM, Vol. 10, pp. 243-244.

IMSL Library 1 Reference Manual. (1973). Houston, Texas.

Jansson, B. (1966). Random Number Generators, Victor Pettersons Bokindustri Atkiebolag, Stockholm.

Jeffreys, H. (1940). "Note on the Behrens-Fisher Formula," Annals of Eugenics, Vol. 10, pp. 48-51.

Lewis, P.A., Goodman, A.S. and Miller, J.M. (1969). "A Pseudo-Random Number Generator for the System/-360," IBM Systems Journal, Vol. 8(2), pp. 136-146.

Mann, H.B. and Wald, A. (1942). "On the Choice of the Number of Class Intervals in the Application of the Chi Square Test," Annals of Mathematical Statistics, Vol. 13, pp. 306-317.

Massey, F.J. Jr. (1951). "The Kolmogorov-Smirnov Test for Goodness of Fit," Journal of the American Statistical Association, Vol. 46, pp. 68-77.

McCullough, R.S., Gurland, J. and Rosenberg, L. (1960). "Small Sample Behavior of Certain Tests of the Hypothesis of Equal Means Under Various Heterogeneity," Biometrika, Vol. 47, pp. 345-353.

Mehta, J.S. and Srinivasan, R. (1970). "On the Behrens-Fisher Problem," Biometrika, Vol. 57, pp. 649-655.

Neyman, J. (1941). "Fiducial Argument and the Theory of Confidence Intervals," Biometrika, Vol. 32, pp. 128-150.

Pearson, E.S. and Hartley, H.O. (1954). Biometrika Tables for Statisticians, Cambridge University Press.

Scheffé, H. (1943). "On Solutions of the Behrens-Fisher Problem, Based on the t Distribution," Annals of Mathematical Statistics, Vol. 14, pp. 35-44.

Scheffé, H. (1970). "Practical Solutions of the Behrens-Fisher Problem," Journal of the American Statistical Association, Vol. 65, pp. 1501-1508.

Seraphin, D.S. (1969). "A Fast Random Number Generator for IBM 360," Communications of the ACM, Vol. 12, pp. 695.

Sukhatme, P.V. (1938). "On Fisher and Behrens' Test of Significance for the Difference in Means of Two Normal Samples," Sankhya, Vol. 4, pp. 39-48.

Trickett, W.H. and Welch, B.L. (1954). "On the Comparison of Two Means: Further Discussion of Iterative Methods for Calculating Tables," Biometrika, Vol. 41, pp. 361-374.

Trickett, W.H., Welch, B.L. and James, G.S. (1956). "Further Critical Values for Two-Means Problem," Biometrika, Vol. 43, pp. 203-205.

Wang, Y.Y. (1971). "Probabilities of the Type 1 Error of the Welch Test for the Behrens-Fisher Problem," Journal of the American Statistical Association, Vol. 66, pp. 605-608.

Welch, B.L. (1938). "The Significance of the Difference Between Two Means When the Population Variances are Unequal," Biometrika, Vol. 29, pp. 350-362.

Welch, B.L. (1947). "The Generalization of Student's Problem When Several Population Variances are Involved," Biometrika, Vol. 34, pp. 28-35.

Welch, B.L. (1956). "Notes on Some Criticisms Made by Sir Ronald Fisher," Journal of Royal Statistical Society B, Vol. 18, pp. 297-302.

West, E.N. (1967). Monte Carlo Investigation of the Behrens-Fisher Problem, Master's Thesis, Iowa State University, Ames, Iowa.

Wilks, S.S. (1940). "On the Problem of Two Samples From Normal Populations With Unequal Variances," Annals of Mathematical Statistics, Vol. 11, pp. 475-476.

Wilson, E.B. and Hilferty, M.M. (1931). "The Distribution of Chi Square," Proceedings National Academy of Science, U.S.A., Vol. 17, pp. 684-688.

Yates, F. (1939). "An Apparent Inconsistency Arising From Tests of Significance Based on Fiducial Distributions of Unknown Parameters," Proceedings of Cambridge Philosophical Society, Vol. 35, pp. 579-591.

Yates, F. (1964). "Fiducial Probability, Recognizable Subsets, and Behrens' Test," Biometrics, Vol. 20, pp. 343-360.